Bio-Inspired Computer Vision: Towards a Synergistic Approach of Artificial and Biological Vision

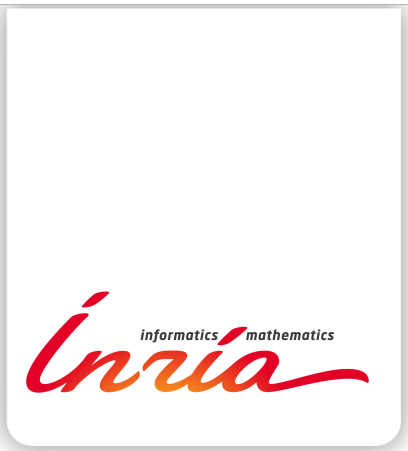
N V Kartheek Medathati, Heiko Neumann, Guillaume S. Masson, Pierre Kornprobst

Bio-Inspired Computer Vision: Towards a Synergistic Approach of Artificial and Biological Vision

N. V. Kartheek Medathati, Heiko Neumann, Guillaume S. Masson, Pierre Kornprobst

RESEARCH REPORT
N° 8698
April 2016
Project-Team Biovision

ISSN 0249-6399 ISRN INRIA/RR--8698--FR+ENG



Bio-Inspired Computer Vision: Towards a Synergistic Approach of Artificial and Biological Vision

N. V. Kartheek Medathati*, Heiko Neumann†, Guillaume S. Masson‡, Pierre Kornprobst*

Project-Team Biovision

Research Report n° 8698 — April 2016 — 71 pages

* Inria Sophia Antipolis M[Pleaseinsert\PrerenderUnicode{Ãľ}intopreamble]diterran[Pleaseinsert\PrerenderUnicode{Ãľ}intopreamble]e, Biovision team, France

† Ulm University, Ulm, Germany

‡ Institut de Neurosciences de la Timone, CNRS & Aix-Marseille Universit[Pleaseinsert\PrerenderUnicode{Ãľ}intopreamble], France

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Abstract: Studies in biological vision have always been a great source of inspiration for design of computer vision algorithms. In the past, several successful methods were designed with varying degrees of correspondence with biological vision studies, ranging from purely functional inspiration to methods that utilise models that were primarily developed for explaining biological observations. Even though it seems well recognised that computational models of biological vision can help in design of computer vision algorithms, it is a non-trivial exercise for a computer vision researcher to mine relevant information from biological vision literature as very few studies in biology are organised at a task level. In this paper we aim to bridge this gap by providing a computer vision task centric presentation of models primarily originating in biological vision studies. Not only do we revisit some of the main features of biological vision and discuss the foundations of existing computational studies modelling biological vision, but also we consider three classical computer vision tasks from a biological perspective: image sensing, segmentation and optical flow. Using this task-centric approach, we discuss well-known biological functional principles and compare them with approaches taken by computer vision. Based on this comparative analysis of computer and biological vision, we present some recent models in biological vision and highlight a few models that we think are promising for future investigations in computer vision. To this extent, this paper provides new insights and a starting point for investigators interested in the design of biology-based computer vision algorithms and pave a way for much needed interaction between the two communities leading to the development of synergistic models of artificial and biological vision.

Key-words: Canonical computations, event based processing, dynamic sensors, multiplexed representation, population coding, soft selectivity, feedback, lateral interactions, form-motion interactions

Vision Bio-Inspirée: Vers une Approche Synergique de la Vision Artificielle et Biologique

Résumé : Les études sur la vision biologique ont toujours été une grande source d'inspiration pour la conception d'algorithmes de vision par ordinateur. Dans le passé, plusieurs méthodes ont été conçues avec succès, avec des degrés variables de correspondance avec les études de la vision biologique, allant de l'inspiration purement fonctionnelle à des procédés qui utilisent des modèles développés principalement pour comprendre les observations biologiques. Même s'il semble bien reconnu que les modèles inspirés du cortex visuel peuvent aider dans la conception d'algorithmes de vision par ordinateur, un exercice non trivial pour un chercheur en vision par ordinateur est de savoir extraire les informations pertinentes de la littérature biologique qui ne s'intéresse que très rarement à la résolution de tâche. Ceci a conduit à un élargissement du fossé entre les recherches menées en vision biologique et en vision par ordinateur. Dans cet article, nous visons à combler cette lacune en procédant à une présentation de la littérature récente en vision biologique orientée vers la résolution de tâches et en fournissant les pointeurs sur des découvertes récentes décrivant les processus sous-jacents. Non seulement nous revisitons certaines des principales caractéristiques de la vision biologique et discutons du fondements des études computationelles modélisant la vision biologique, mais aussi nous revisitons trois tâches classiques en vision par ordinateur avec un point de vue biologique: l'acquisition d'images, la segmentation et le flot optique. En utilisant cette approche orientée vers la résolution des tâches, nous discutons des principes fonctionnels biologiques connus pour les comparer avec les approches proposées en vision par ordinateur. Sur la base de cette analyse comparative entre vision biologique et artificielle, nous présentons des approches prometteuses récentes en modélisation de la vision biologique et nous soulignons des idées nouvelles qui nous paraissent prometteuses pour les recherches futures en vision par ordinateur. En ce sens, ce papier offre de nouvelles perspectives pour la conception d'algorithmes de vision inspirés de la biologie et il ouvre une voie à une interaction indispensable entre les modélisateurs des deux communautés.

Mots-clés : Calculs canoniques, calcul évênementiels, capteurs dynamiques, représentation multiplexée, codage en population, sélectivité souple, slVision bio-inspirée, acquisition, segmentation, flot optique, retina, voie dorsale, voir ventrale

Contents

1 Introduction **5**

2 Deep cortical hierarchies? **7**
2.1 The classical view of biological vision 7
2.2 Going beyond the hierarchical feedforward view 8

3 Computational studies of biological vision **15**
3.1 The Marr's three levels of analysis 15
3.2 From circuits to behaviours 15
3.3 Neural constraints for functional tasks 16
3.4 Matching connectivity rules with computational problems 17
3.5 Testing biologically-inspired models against both natural and computer vision . . 19
3.6 Task-based versus general purpose vision systems 19

4 Solving vision tasks with a biological perspective **21**
4.1 Sensing 21
4.2 Segmentation and figure-ground segregation 26
4.3 Optical flow 33

5 Discussion **40**
5.1 Structural principles that relate to function 40
5.2 Data encoding and representation 42
5.3 Psychophysics and human perceptual performance data 43
5.4 Computational models of cortical processing 44

6 Conclusion **47**

1 Introduction

Biological vision systems are remarkable at extracting and analysing the essential information for vital functional needs such as navigating through complex environments, finding food or escaping from a danger. It is remarkable that biological visual systems perform all these tasks with both high sensitivity and strong reliability given the fact that natural images are highly noisy, cluttered, highly variable and ambiguous. Still, even simple biological systems can efficiently and quickly solve most of the difficult computational problems that are still challenging for artificial systems such as scene segmentation, local and global optical flow computation, 3D perception or extracting the meaning of complex objects or movements. All these aspects have been intensively investigated in human psychophysics and the neuronal underpinnings of visual performance have been scrutinised over a wide range of temporal and spatial scales, from single cell to large cortical networks so that visual systems are certainly the best-known of all neural systems (see [59] for an encyclopaedic review). As a consequence, biological visual computations are certainly the most understood of all cognitive neural systems.

It would seem natural that biological and computer vision research would interact continuously since they target the same goals at task level: extracting and representing meaningful visual information for making actions. Sadly, the strength of these interactions has remained weak since the pioneering work of David Marr [202] and colleagues who attempted to marry the fields of neurobiology, visual psychophysics and computer vision. The unifying idea presented in his influential book entitled Vision was to articulate these fields around computational problems faced by both biological and artificial systems rather than on their implementation. Despite these efforts, the two research fields have however largely drifted apart, partly because of several technical obstacles that obstructed this interdisciplinary agenda for decades, such as the limited capacity of the experimental tools used to probe visual information processing or the limited computational power available for simulations.

With the advent of new experimental and analysis techniques significant amount of progress has been made towards overcoming these technical obstacles. A new wealth of multiple scales functional analysis and connectomics information is emerging in brain sciences, and it is encouraging to note that studies of visual systems are upfront on this fast move [81]. For instance, it is now possible to identify selective neuronal populations and dissect out their circuitry at synaptic level by combining functional and structural imaging. The first series of studies applying such techniques have focused on understanding visual circuits, at both retinal [128] and cortical [31] levels. At a wider scale, a quantitative description of the connectivity patterns between cortical areas is now becoming available and, here again the study of visual cortical networks is pioneering [200]. A first direct consequence is that detailed large scales models of visual networks are now available to study the neurobiological underpinnings of information processing at multiple temporal and spatial scales [160, 270, 60]. With the emergence of international research initiatives (e.g., the BRAIN and HBP projects, the Allen Institute Atlas), we are certainly at the first steps of a major revolution in brain sciences. At the same time, recent advances in computer architectures make it now possible to simulate large-scale models, something that was not even possible to dream of a few years ago. For example, the advent of multi-core architectures [82], parallel computing on clusters [263], GPU computing [261] and availability of neuromorphic hardware [333] promises to facilitate the exploration of truly bio-inspired vision systems [218]. However, these technological advancements in both computer and brain sciences call for a strong push in theoretical studies. The theoretical difficulties encountered by each field call for a new, interdisciplinary approach for understanding how we process, represent and use visual information. For instance, it is still unclear how the dense network of cortical areas fully analyses the structure of the external world and part of the problem may come from using a bad range of

framing questions about mid-level and high-level vision [67, 170, 120]. In short, we cannot see the forest (representing the external world) for the trees (e.g., solving face and object recognition) and reconciling biological and computer vision is a timely joint-venture for solving these challenges.

The goal of this paper is to advocate how novel computer vision approaches could be developed from these biological insights. It is a manifesto for developing and scaling up models rooted in experimental biology (neurophysiology, psychophysics, etc.) leading to an exciting synergy between studies in computer vision and biological vision. Our conviction is that the exploding knowledge about biological vision, the new simulation technologies and the identification of some ill-posed problems have reached a critical point that will nurture a new departure for a fruitful interdisciplinary endeavour. The resurgence of interest in biological vision as a rich source for designing principles for computer vision is evidenced by recent books [258, 98, 129, 266, 70, 189] and survey papers [346, 68]. However, we feel that these studies were more focused on computational neuroscience rather than computer vision and, second remain largely influenced by the hierarchical feedforward approach, thus ignoring the rich dynamics of feedback and lateral interactions.

This article is organised as follows. In Sec. 2, we revisit the classical view of the brain as a hierarchical feedforward system [169]. We point out its limitations and portray a modern perspective of the organisation of the primate visual system and its multiple spatial and temporal anatomical and functional scales. In Sec. 3, we appraise the different current computational and theoretical frameworks used to study biological vision and re-emphasise the importance of putting the task solving approach as the main motivation to look into biology. In order to relate studies in biological vision to computer vision, we focus in Sec. 4 on three archetypal tasks: sensing, segmentation and motion estimation. These three tasks are illustrative because they have similar basic-level representations in biological and artificial vision. However, the role of the intricate, recurrent neuronal architecture in figuring out neural solutions must be re-evaluated in the light of recent empirical advances. For each task, we will start by highlighting some of these recently-identified biological mechanisms that can inspire computer vision. We will give a structural view of these mechanisms, relate these structural principles to prototypical models from both biological and computer vision and, finally we will detail potential insights and perspectives for rooting new approaches on the strength of both fields. Finally, based on the prototypical tasks reviewed throughout this article, we will propose in Sec. 5, three ways to identify which studies from biological vision could be leveraged to advance computer vision algorithms.

2 Deep cortical hierarchies?

2.1 The classical view of biological vision

The classical view of biological visual processing that has been conveyed to the computer vision community from visual neurosciences is that of an ensemble of deep cortical hierarchies (see [169] for a recent example). Interestingly, this computational idea was proposed in computer vision by David Marr [202] even before its anatomical hierarchy was fully detailed in different species. Nowadays, there is a general agreement about this hierarchical organisation and its division into parallel streams in human and non-human primates, as supported by a large body of anatomical and physiological evidences (see [356, 358, 200] for reviews). Fig. 1(a)–(b) illustrates this classical view where information flows from the retina to the primary visual cortex (area V1) through two parallel retino-geniculo-cortical pathways. The magnocellular (M) pathway conveys coarse, luminance-based spatial inputs with a strong temporal sensitivity towards Layer 4Cα of area V1 where a characteristic population of cells, called stellate neurons, immediately transmit the information to higher cortical areas involved in motion and space processing. A slower, parvocellular (P) pathway conveys retino-thalamo-cortical inputs with high spatial resolution but low temporal sensitivity, entering area V1 through the layer 4Cβ. Such color-sensitive input flows more slowly within the different layers of V1 and then to cortical area V2 and a network of cortical areas involved in form processing. The existence of these two parallel retino-thalamo-cortical pathways resonated with neuropsychological studies investigating the effects of parietal and temporal cortex lesions [355], leading to the popular, but highly schematic, two visual systems theory [355, 356, 220] in which a dorsal stream is specialised in motion perception and the analysis of the spatial structure of the visual scene whereas a ventral stream is dedicated to form perception, including object and face recognition.

At the computational level, the deep hierarchies concept was reinforced by the linear systems approach used to model low-level visual processing. As illustrated in Fig. 1(c), neurons in the primary visual system have small receptive fields, paving a high resolution retinotopic map. The spatiotemporal structure of each receptive field corresponds to a processing unit that locally filters a given property of the image. In V1, low-level features such as orientation, direction, color or disparity are encoded in different sub-populations forming a sparse and overcomplete representation of local feature dimensions. These representations feed several, parallel cascades of converging influences so that, as one moves along the hierarchy, receptive fields become larger and larger and encode for features of increasing complexities and conjunctions thereof (see [76, 293] for reviews). For instance, along the motion pathway, V1 neurons are weakly direction-selective but converge onto the medio-temporal (MT) area where cells can precisely encode direction and speed in a form-independent manner. These cells project to neurons in the median superior temporal (MST) area where receptive fields cover a much larger portion of the visual field and encode basic optic flow patterns such as rotation, translation or expansion. More complex flow fields can be decoded by parietal neurons when integrating these informations and be integrated with extra-retinal signals about eye movements or self-motion [39, 244]. The same logic flows along the form pathway, where V1 neurons encode the orientation of local edges. Through a cascade of convergence, units with receptive fields sensitive to more and more complex geometrical features are generated so that neurons in the infero-temporal (IT) area are able to encode objects or face in a viewpoint invariant manner (see Fig. 1(c)).

Object recognition is a prototypical example where the canonical view of hierarchical feedforward processing nearly perfectly integrates anatomical, physiological and computational knowledges. This synergy has resulted in realistic, computational models of receptive fields where converging outputs from linear filters are nonlinearly combined from one step to the subsequent

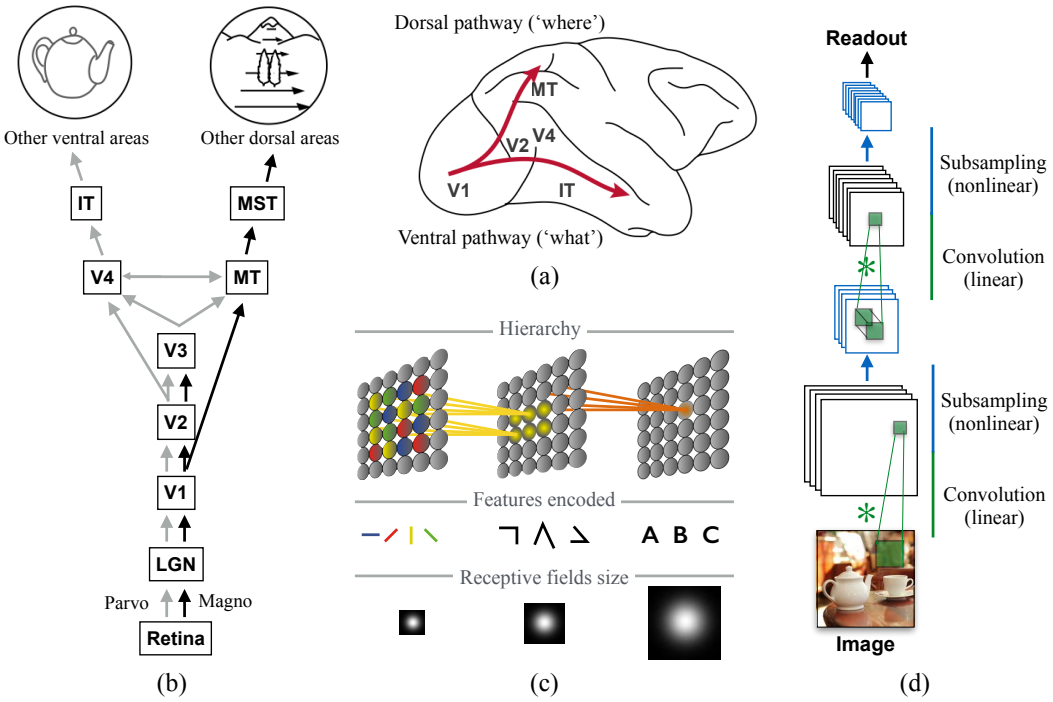


Figure 1: The classical view of hierarchical feedforward processing. (a) The two visual pathways theory states that primate visual cortex can be split between dorsal and ventral streams originating from the primary visual cortex (V1). The dorsal pathway runs towards the parietal cortex, through motion areas MT and MST. The ventral pathway propagates through area V4 all along the temporal cortex, reaching area IT. (b) These ventral and dorsal pathways are fed by parallel retino-thalamo-cortical inputs to V1, known as the Magno (M) and Parvocellular pathways (P). (c) The hierarchy consists in a cascade of neurons encoding more and more complex features through convergent information. By consequence, their receptive field integrate visual information over larger and larger receptive fields. (d) Illustration of a machine learning algorithm for, e.g., object recognition, following the same hierarchical processing where a simple feedforward convolutional network implements two bracketed pairs of convolution operator followed by a pooling layer (adapted from [68]).

one [53, 232]. It has also inspired feedforward models working at task levels for object categorisation [310, 309] as illustrated in Fig. 1(d), prominent machine learning solutions for object recognition follow the same feedforward, hierarchical architecture where linear and nonlinear stages are cascaded between multiple layers representing more and more complex features [135, 68].

2.2 Going beyond the hierarchical feedforward view

Despite its success in explaining some basic aspects of human perception such as object recognition, the hierarchical feedforward theory remains highly schematic. Many aspects of biological visual processing, from anatomy to behaviour, do not fit in this cartoon-like framing. Important aspects of human perception such as detail preservation, multi-stability, active vision and space perception for example cannot be adequately explained by a hierarchical cascade of expert cells. Furthermore, taking into account high-level cognitive skills such as top-down attention, visual

cognition or concepts representation needs to reconsider this deep hierarchies. In particular, the dynamics of neural processing is much more complex than the hierarchical feedforward abstraction and very important connectivity patterns such as lateral and recurrent interactions must be taken into account to overcome several pitfalls in understanding and modelling biological vision. In this section, we highlight some of these key novel features that should greatly influence computational models of visual processing. We also believe that identifying some of these problems could help in reunifying natural and artificial vision and addressing more challenging questions as needed for building adaptive and versatile artificial systems which are deeply bio-inspired.

Vision processing starts at the retina and the lateral geniculate nucleus (LGN) levels. Although this may sound obvious, the role played by these two structures seems largely underestimated. Indeed, most current models take images as inputs rather than their retina-LGN transforms. Thus, by ignoring what is being processed at these levels, one could easily miss some key properties to understand what makes the efficiency of biological visual systems. At the retina level, the incoming light is transformed into electrical signals. This transformation was originally described by using the linear systems approach to model the spatio-temporal filtering of retinal images [86]. More recent research has changed this view and several cortex-like computations have been identified in the retina of different vertebrates (see [110, 156] for reviews, and more details in Sec. 4.1). The fact that retinal and cortical levels share similar computational principles, albeit working at different spatial and temporal scales is an important point to consider when designing models of biological vision. Such a change in perspective would have important consequences. For example, rather than considering how cortical circuits achieve high temporal precision of visual processing, one should ask how densely interconnected cortical networks can maintain the high temporal precision of the retinal encoding of static and moving natural images [92], or how miniature eye movements shapes its spatiotemporal structure [299].

Similarly, the LGN and other visual thalamic nuclei (e.g., pulvinar) should no longer be considered as pure relays on the route from retina to cortex. For instance, cat pulvinar neurons exhibit some properties classically attributed to cortical cells, as such pattern motion selectivity [217]. Strong centre-surround interactions have been shown in monkeys LGN neurons and these interactions are under the control of feedback cortico-thalamic connections [152]. These strong cortico-geniculate feedback connections might explain why parallel retino-thalamo-cortical pathways are highly adaptive, dynamical systems [229, 72, 41]. In line with the computational constraints discussed before, both centre-surround interactions and feedback modulation can shape the dynamical properties of cortical inputs, maintaining the temporal precision of thalamic firing patterns during natural vision [6].

Overall, recent sub-cortical studies give us three main insights. First, we should not oversimplify the amount of processing done before visual inputs reach the cortex and we must instead consider that the retinal code is already highly structured, sparse and precise. Thus, we should consider how cortex takes advantage of these properties when processing naturalistic images. Second, some of the computational and mechanistic rules designed for predictive-coding or feature extraction can be much more generic than previously thought and the retina-LGN processing hierarchy may become again a rich source of inspiration for computer vision. Third, the exact implementation (what is being done and where) may be not so important as it varies from one species to another but the cascade of basic computational steps may be an important principle to retain from biological vision.

Functional and anatomical hierarchies are not always identical. The deep cortical hierarchy depicted in Fig. 1(b) is primarily based on gross anatomical connectivity rules [383]. Its functional counterpart is the increasing complexity of local processing and information content

of expert cells as we go deeper along the anatomical hierarchy. There is however a flaw in attributing the functional hierarchy directly to its anatomical counterpart. The complexity of visual processing does increase from striate to extra-striate and associative cortices, but this is not attributable only to feedforward convergence. A quick glance at the actual cortical connectivity pattern in non-human primates would be sufficient to eradicate this textbook view of how the visual brain works [126, 200].

For example, a classical view is that the primary visual cortex represents luminance-based edges whereas higher-order image properties such as illusory contours are encoded at the next processing stages along the ventral path (e.g., areas V2 and V4) [255]. Recent studies have shown however that illusory contours, as well as border ownerships can also be represented in macaque area V1 [385, 180]. Moreover, multiple binocular and monocular depth cues can be used to reconstruct occluded surfaces in area V1 [330]. Thus, the hierarchy of shape representation appears nowadays more opaque than previously thought [127] and many evidences indicate that the intricate connectivity within and between early visual areas is decisive for of the emergence of figure-ground segmentation and proto-objects representations [259, 364]. Another strong example is visual motion processing. The classical feedforward framework proposes that MT cells (and not V1 cells) are true speed-tuned units. It has been thought for decades that V1 cells cannot encode the speed of a moving pattern independently of its spatiotemporal frequencies content [288]. However, recent studies have shown that there are V1 complex cells which are speed tuned [275]. The differences between V1 and MT regarding speed coding are more consistent with a distributed representation where slow speeds are represented in V1 and high speeds in area MT rather than a pure, serial processing. Decoding visual motion information at multiple scales for elaborating a coherent motion percept must therefore imply a large-scale cortical network of densely recurrently interconnected areas. Such network can extend to cortical areas along the ventral stream in order to integrate together form and complex global motion inputs [386, 124]. One final example concerns the temporal dynamics of visual processing. The temporal hierarchy is not a carbon copy of the anatomical hierarchy depicted by Felleman and Van Essen. The onset of a visual stimulus triggers fast and slow waves of activation travelling throughout the different cortical areas. The fast activation in particular by-passes several major steps along both dorsal and ventral pathways to reach frontal areas even before area V2 is fully activated (for a review, see [175]). Moreover, different time scales of visual processing emerge from both the feedforward hierarchy of cortical areas but also from the long-range connectivity motifs and the dense recurrent connectivity of local sub-networks [60]. Such rich repertoire of temporal time windows, ranging from fast, transient responses in primary visual cortex to persistent activity in association areas, is critical for implementing a series of complex cognitive tasks from low-level processing to decision-making.

These three different examples highlight the fact that a more complex view of the functional hierarchy is emerging. The dynamics of biological vision results from the interactions between different cortical streams operating at different speeds but also relies on a dense network of intra-cortical and inter-cortical (e.g., feedback) connections. Designing better vision algorithms could be inspired by this recurrent architecture where different spatial and temporal scales can be mixed to represent visual motion or complex patterns with both high reliability and high resolution.

Dorsal/ventral separation is an over-simplification. A strong limitation of grounding a theoretical framework of sensory processing upon anatomical data is that the complexity of connectivity patterns must lead to undesired simplifications in order to build a coherent view of the system. Moreover, it escapes the complexity of the dynamical functional interactions between areas or cognitive sub-networks. A good example of such bias is the classical dorsal/ventral sep-

aration. First, interactions between parallel streams can be tracked down to the primary visual cortex where a detailed analysis of the layer 4 connectivity have shown that both Magno and Parvocellular signals can be intermixed and propagated to areas V2 and V3 and, therefore the subsequent ventral stream [378]. Such a mixing of M- and P-like signals could explain why fast and coarse visual signals can rapidly tune the most ventral areas along the temporal cortex and therefore shape face recognition mechanisms [105]. Second, motion psychophysics has demonstrated a strong influence of form signals onto local motion analysis and motion integration [210]. These interactions have been shown to occur at different levels of the two parallel hierarchies, from primary visual cortex to the superior temporal sulcus and the parietal cortex [244]. These interactions provide many computational advantages used by the visual motion system to resolve motion ambiguities, interpolate occluded information, segment the optical flow or recover the 3D structure of objects. Third, there are strong interactions between color and motion information, through mutual interactions between cortical areas V4 and MT [334]. It is interesting to note that these two particular areas were previously attributed to the ventral and dorsal pathways, respectively [191, 76]. Such strict dichotomy is outdated as both V4 and MT areas interact to extract and mix these two dimensions of visual information.

These interactions are only a few examples to be mentioned here to highlight the needs of a more realistic and dynamical model of biological visual processing. If the coarse division between ventral and dorsal streams remains valid, a closer look at these functional interactions highlight the existence of multiple links, occurring at many levels along the hierarchy. Each stream is traversed by successive waves of fast/coarse and slow/precise signals so that visual representations are gradually shaped [290]. It is now timely to consider the intricate networks of intra and inter-cortical interactions to capture the dynamics of biological vision. Clearly, a new theoretical perspective on the cortical functional architecture would be highly beneficial to both biological and artificial vision research.

A hierarchy embedded within a dynamical recurrent system. We have already mentioned that spatial and temporal hierarchies do not necessarily coincide as information flows can bypass some cortical areas through fast cortico-cortical connections. This observation led to the idea that fast inputs carried by the Magnocellular stream can travel quickly across the cortical networks to shape each processing stage before it is reached by the fine-grain information carried by the Parvocellular retino-thalamo-cortical pathway. Such dynamics are consistent with the feedforward deep hierarchy and are used by several computational models to explain fast, automatic pattern recognition [298, 337].

Several other properties of visual processing are more difficult to reconcile with the feedforward hierarchy. Visual scenes are crowded and it is not possible to process every of its details, Moreover, visual inputs are often highly ambiguous and can lead to different interpretations, as evidenced by perceptual multi-stability. Several studies have proposed that the highly recurrent connectivity motif of the primate visual system plays a crucial role in these processing. At the theoretical level, several authors recently resurrected the idea of a “reversed hierarchy" where high-level signals are back-propagated to the earliest visual areas in order to link low-level visual processing, high resolution representation and cognitive information [48, 136, 3, 120]. Interestingly, this idea was originally proposed more than three decades before by Peter Milner in the context of visual shape recognition [221] and had then quickly diffused to the computer vision research leading to novel algorithms for top-down modulation, attention and scene parsing (e.g., [99, 343, 345]). At the computational level, in [179] the authors reconsidered the hierarchical framework by proposing that concatenated feedforward/feedback loops in the cortex could serve to integrate top-down prior knowledge with bottom-up observations. This architecture generates a cascade of optimal inference along the hierarchy [293, 179, 298, 337]. Several computational

models have used such recurrent computation for surface motion integration [21, 340, 252], contour tracing [44] or figure-ground segmentation [294].

Empirical evidence for a role of feedback has long been difficult to gather in support to these theories. It was thus difficult to identify the constraints of top-down modulations that are known to play a major role in the processing of complex visual inputs, through selective attention, prior knowledge or action-related internal signals. However, new experimental approaches begin to give a better picture of their role and their dynamics. For instance, selective inactivation studies have begun to dissect the role of feedback signals in context-modulation of primate LGN and V1 neurons [72]. The emergence of genetically-encoded optogenetic probes targeting the feedback pathways in mice cortex opens a new era of intense research about the role of feedforward and feedback circuits [195, 145]. Overall, early visual processing appears now to be strongly influenced by different top-down signals about attention, working memory or even reward mechanisms, just to mention. These new empirical studies pave the way for a more realistic perspective on visual perception where both sensory inputs and brain states must be taken into account when, for example, modelling figure-ground segmentation, object segregation and target selection (see [175, 327, 153] for recent reviews).

The role of attention is illustrative of this recent trend. Mechanisms of bottom-up and top-modulation attentional modulations in primates have been largely investigated over the last three decades. Spatial and feature-based attentional signals have been shown to selectively modulate the sensitivity of visual responses even in the earliest visual areas [226, 287]. These works have been a vivid source of inspiration for computer vision in searching for a solution to the problems of feature selection, information routing and task-specific attentional bias (see [146, 344]), as illustrated for instance by the Selective Tuning algorithm of Tsotsos and collaborators [345]. More recent work in non-human primates has shown that attention can also affect the tuning of individual neurons [144]. It also becomes evident that one needs to consider the effects of attention on population dynamics and the efficiency of neural coding (e.g., by decreasing noise correlation [64]). Intensive empirical work is now targeting the respective contributions of the frontal (e.g., task-dependency) and parietal (e.g., saliency maps) networks in the control of attention and its coupling with other cognitive processes such as reward learning or working memory (see [50] for a recent review). These empirical studies led to several computational models of attention (see [344, 347, 52] for recent reviews) based on generic computations (e.g., divisive normalisation [286], synchrony [97] or feedback-feedforward interactions [158]). Nowadays, attention appears to be a highly dynamical, rapidly changing processing that recruits a highly flexible cortical network depending on behavioural demands and in strong interactions with other cognitive networks.

The role of lateral connectivity in information diffusion. The processing of a local feature is always influenced by its immediate surrounding in the image. Feedback is one potential mechanisms for implementing context-dependent processing but its spatial scale is rather large, corresponding to far-surround modulation [8]. Visual cortical areas, and in particular area V1, are characterised by dense short- and long-range intra-cortical interactions. Short-range connectivities are involved in proximal centre-surround interactions and their dynamics fits with contextual modulation of local visual processing [285]. This connectivity pattern has been overly simplified as overlapping, circular excitatory and inhibitory areas of the non-classical receptive field. In area V1, these sub-populations were described as being tuned for orthogonal orientations corresponding to excitatory input from iso-oriented domains and inhibitory input from cross-oriented ones. In higher areas, similar simple schemes have been proposed, such as the opposite direction tuning of center and surround areas of MT and MST receptive fields [33]. Lastly, these surround inputs have been proposed to implement generic neural computations

such as normalisation or gain control [57].

From the recent literature, a more complex picture of centre-surround interactions has emerged where non-classical receptive fields are highly diverse in terms of shapes or features selectivity [377, 58, 369]. Such diversity would result from complex connectivity patterns where neurons tuned for different features (e.g., orientation, direction, spatial frequency) can be dynamically interconnected. For example, in area V1, the connectivity pattern becomes less and less specific with farther distances from the recording sites. Moreover, far away points in the image can also interact through the long-range interactions which have been demonstrated in area V1 of many species. Horizontal connections extend over millimetres of cortex and propagate activity at a much lower speed than feedforward and feedback connections [48]. The functional role of these long-range connections is still unclear. They most probably support the waves of activity that travel across the V1 cortex either spontaneously or in response to a visual input [303, 228]. They can also implement the spread of cortical activity underlying contrast normalisation [285], the spatial integration of motion and contour signals [285, 107] or the shaping of low-level percepts [147].

A neural code for vision? How is information encoded in neural systems is still highly disputed and an active field of theoretical and empirical research. Once again, visual information processing has been largely used as a seminal framework to decipher the neural coding principles and its application for computer sciences. The earliest studies on neuronal responses to visual stimuli have suggested that information is encoded in the mean firing rate of individual cells and its gradual change with visual input properties. For instance cells in V1 labelled as feature detectors are classified based upon their best response selectivity (stimulus that invokes maximal firing of the neuron) and several non-linear properties such gain control or context modulations which usually varied smoothly with respect to few attributes such as orientation contrast and velocity, leading to the development of tuning curves and receptive field doctrine. Spiking and mean-field models of visual processing are based on these principles.

Aside of from changes in mean firing rates, other interesting features of neural coding is the temporal signature of neural responses and the temporal coherence of activity between ensembles of cells, providing an additional potential dimension for specific linking, or grouping, distant and different features [365, 366, 323]. In networks of coupled neuronal assemblies, associations of related sensory features are found to induce oscillatory activities in a stimulus-induced fashion [79]. The establishment of a temporal coherence has been suggested to solve the so-called binding problem of task-relevant features through synchronization of neuronal discharge patterns in addition to the structural patterns of linking pattern [85]. Such synchronizations might even operate over different areas and therefore seems to support rapid formations of neuronal groups and functional subnetworks and routing signals [97, 50]. However, the view that temporal oscillatory states might define a key element of feature coding and grouping has been challenged by different studies and the exact contribution of these temporal aspects of neural codes is not yet fully elucidated (e.g., [311] for a critical review). By consequences, only a few of bio-inspired and computer vision models rely on the temporal coding of information.

Although discussing the many facets of visual information coding is far beyond the scope of this review, one needs to briefly recap some key properties of neural coding in terms of tuning functions. Representations based on the tuning functions can be basis for the synergistic approach advocated in this article. Neurons are tuned to one or several features, i.e., exhibiting a strong response when stimuli contrains a preferred feature such as local luminance-defined edges or proto-objects and low or no response when such features are absent. As a result, neural feature encoding is sparse, distributed over populations (see [271, 312] and highly reliable [250] at the same time. Moreover, these coding properties emerge from the different connectivity

rules introduced above. The tuning functions of individual cells are very broad such that high behavioural performances observed empirically can be achieved only from some nonlinear or probabilistic decoding of population activities [271]. This could also imply that visual information could be represented within distributed population codes rather than grand-mother cells [273, 181]. Tuning functions are dynamical: they can be sharpened or shifted over time [315]. Neural representation could also be relying on spike timing and the temporal structure of the spiking patterns can carry additional information about the dynamics of transient events [338, 251]. Overall, the visual system appears to use different types of codes, one advantage for representing high-dimension inputs [297].

3 Computational studies of biological vision

3.1 The Marr's three levels of analysis

At conceptual level, much of the current computational understanding of biological vision is based on the influential theoretical framework defined by David Marr [202] and colleagues. Their key message was that complex systems, like brains or computers, must be studied and understood at three levels of description: the computational task carried out by the system resulting in the observable behaviour, the instance of the algorithm used by the system to solve the computational task and the implementation that is emboddied by a given system to execute the algorithm. Once a functional framework is defined, the computational and implementation problems can be distinguished, so that in principle a given solution can be embedded into different biological, or artificial physical systems. This approach has inspired many experimental and theoretical research in the field of vision [111, 133, 74, 264]. The cost of this clear distinction between levels of description is that many of the existing models have only a weak relationship with the actual architecture of the visual system or even with a specific algorithmic strategy used by biological systems. Such dichotomy contrasts with the growing evidence that understanding cortical algorithms and networks are deeply coupled [133]. Human perception would still act as a benchmark or a source of inspiring computational ideas for specific tasks (see [7] for a good example about object recognition). But, the risk of ignoring the structure-function dilemma is that computational principles would drift away from biology, becoming more and more metaphorical as illustrated by the fate of the Gestalt theory. The bio-inspired research stream for both computer vision and robotics aims at reducing this fracture (e.g. [258, 129, 98, 70] for recent reviews).

3.2 From circuits to behaviours

A key milestone in computational neurosciences is to understand how neural circuits lead to animal behaviours. Carandini [55] argued that the gap between circuits and behaviour is too wide without the help of an intermediate level of description, just that of neuronal computation. But how can we escape from the dualism between computational algorithm and implementation as introduced by Marr's approach? The solution depicted in [55] is based on three principles. First, some levels of description might not be useful to understand functional problems. In particular sub cellular and network levels are decoupled. Second, the level of neuronal computation can be divided into building blocks forming a core set of canonical neural computations such as linear filtering, divisive normalisation, recurrent amplification, coincidence detection, cognitive maps and so on. These standard neural computations are widespread across sensory systems [95]. Third, these canonical computations occur in the activity of individual neurons and especially of population of neurons. In many instances, they can be related to stereotyped circuits such as feedforward inhibition, recurrent excitation-inhibition or the canonical cortical microcircuit for signal amplification (see [316] for a series of reviews). Thus, understanding the computations carried out at the level of individual neurons and neural populations would be the key for unlocking the algorithmic strategies used by neural systems. This solution appears to be essential to capture both the dynamics and the versatility of biological vision. With such a perspective, computational vision would regain its critical role when mapping circuits to behaviours and could rejuvenate the interest in the field of computer vision not only by highlighting the limits of existing algorithms or hardware but also by providing new ideas. At this cost, visual and computational neurosciences would be again a source of inspiration for computer vision. To illustrate this joint venture, Figure 2 illustrates the relationships between the different functional and anatomical scales of cortical processing and their mapping with the three computational problems encountered with designing any artificial systems:how, what and why.

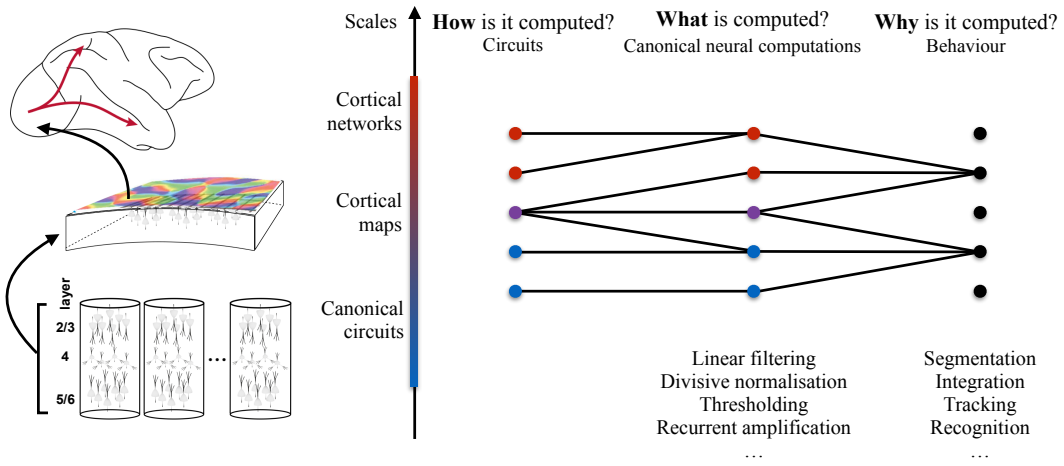


Figure 2: Between circuits and behaviour: rejuvenating the Marr approach. The nervous system can be described at different scales of organisation that can be mapped onto three computational problems: how, what and why. All three aspects involve a theoretical description rooted on anatomical, physiological and behaviour data. These different levels are organised around computational blocks that can be combined to solve a particular task.

3.3 Neural constraints for functional tasks

Biological systems exist to solve functional tasks so that an organism can survive. Considering the existing constraints, many biologists consider the brain as a "bag of tricks that passed evolutionary selection", even though some tricks can be usable in different systems or contexts. This biological perspective highlights the fact that understanding biological systems is tightly related to understanding the functional importance of the task at hands. For example, there is in the mouse retina a cell type able to detect small moving objects in the presence of a featureless or stationary background. These neurons could serve as elementary detectors of potential predators arriving from the sky [384]. In the same vein, it has been recently found that output of retinal direction-selective cells are kept separated from the other retino-thalamo-cortical pathways to directly influence specific target neurons in mouse V1 [71]. These two very specific mechanisms illustrate how evolution can shape nervous systems. Computation and architecture are intrinsically coupled to find an optimal solution. This could be taken as an argument for ignoring neural implementations when building generic artificial systems. However, there are also evidence that evolution has selected neural microcircuits implementing generic computations such as divisive normalisation. These neural computations have been shown to play a key role in the emergence of low-level neuronal selectivities. For example divisive normalisation has been a powerful explanation for many aspects of visual perception, from low-level gain control or attention [286, 57]. The role of feedforward-feedback connectivity rules of canonical microcircuits in predictive coding have been also identified [19] and applied in the context of visual motion processing [78]. These examples are extrema lying on the continuum of biological structure-function solutions, from the more specific to the more generic. This diversity stresses the needs to clarify the functional context of the different computational rules and their performance dynamics so that fruitful comparisons can be made between living and artificial systems. This can lead to a clarification about which knowledge from biology is useful for computer vision.

Lastly, these computational building blocks are embedded into a living organism and low-to-

high vision levels are constantly interacting with many other aspects of animal cognition [362]. For example, the way an object is examined (i.e., the way its image is processed) depends on its behavioural context, whether it is going to be manipulated or only scrutinised to identify it. A single face can be analysed in different ways depending upon the social or emotional context. Thus, we must consider these contextual influence of "why" a task is being carried out when integrating information (and data) from biology [372]. All these above observations stress the difficulty of understanding biological vision as an highly adapted, plastic and versatile cognitive system where circuits and computation are like Janus face. However, as described above for recurrent systems, understanding the neural dynamics of versatile top-down modulation can inspired artificial systems about how different belief states can be integrated together within the low-level visual representations.

3.4 Matching connectivity rules with computational problems

In Sec. 2, we have given a brief glimpse of the enormous literature on the intricate networks underlying biological vision. Focusing on primate low-level vision, we have illustrated both the richness, the spatial and temporal heterogeneity and the versatility of these connections. We illustrate them in Fig. 3 for a simple case, the segmentation of two moving surfaces. Figure 3(a) sketches the main cortical stages needed for a minimal model of surface segmentation [244, 340]. Local visual information is transmitted upstream through the retinotopicaly-organized feedforward projections. In the classical scheme, V1 is seen as a router filtering and sending the relevant information along the ventral (V2, V4) or dorsal (MT, MST) pathways [169]. We discussed above how information flows also backward within each pathway as well as across pathways, as illustrated by connections between V2/V4 and MT in Fig. 3) [201]. One consequence of these cross-over is that MT neurons are able to use both motion and color information [334]. We have also highlighted that area V1 endorses a more active role where the thalamo-cortical feedforward inputs and the multiple feedback signals interact to implement contextual modulations over different spatial and temporal scales using generic neural computations such surround suppression, spatio-temporal normalisation and input selection. These local computations are modulated by short and long-range intra-cortical interactions such as visual features located far from the non-classical receptive field (or along a trajectory) can influence them [9]. Each cortical stage implements these interactions although with different spatial and temporal windows and through different visual feature dimensions. In Fig. 3, these interactions are illustrated within two (orientation and position) of the many cortical maps founds in both primary and extra-striate visual areas. At the single neuron level, these intricate networks result in a large diversity of receptive field structures and in complex, dynamical non-linearities. It is now possible to collect physiological signatures of these networks at multiple scales, from single neurons to local networks and networks-of-networks such that connectivity patterns can be dissected out. In the near future, it will become possible to manipulate specific cell subtype and therefore change the functional role and the weight of these different connectivities.

How these connectivity patterns would relate to information processing? In Fig. 3(b) as an example, we sketch the key computational steps underlying moving surface segmentation [38]. Traditionally, each computational step has been attributed to a particular area and to a specific type of receptive fields. For instance, local motion computation is done at the level of the small receptive fields of V1 neurons. Motion boundary detectors have been found in area V2 while different subpopulation of MT and MST neurons are responsible for motion integration at multiple scales (see Sec. 4.3 for references). However, each of these receptive field types are highly context-dependent, as expected from the dense interactions between all these areas. Matching the complex connectivity patterns illustrated in Fig. 3(a) with the computational dynamics

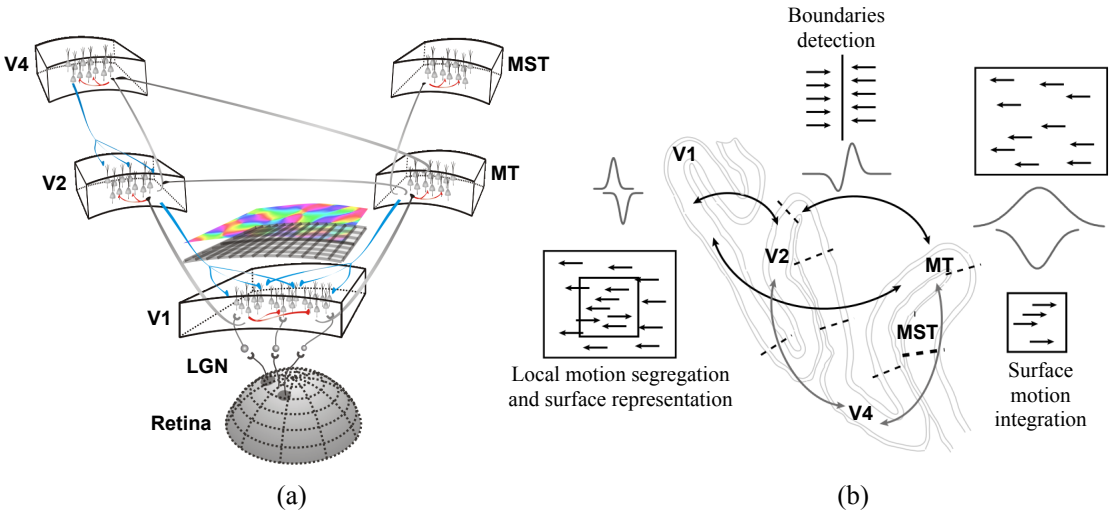


Figure 3: Matching multi-scale connectivity rules and computational problems for the segmentation of two moving surfaces. (a) A schematic view of the early visual stages with their different connectivity patterns: feedforward (grey), feedback (blue) and lateral (red). (b) A sketch of the problem of moving surface segmentation and its potential implementation in the primate visual cortex. The key processing elements are illustrated as computational problems (e.g., local segregation, surface cues, motion boundaries, motion integration) and corresponding receptive field structures. These receptive fields are highly adaptive and reconfigurable, thanks to the dense interconnections between the different stages/areas

illustrated in Fig. 3(b) is one of the major challenges in computational neurosciences [95]. But it could also be a fruitful source of inspiration for computer vision if we were able to draw the rules and numbers by which the visual system is organised at different scales. So far, only a few computational studies have taken into account this richness and its ability to adaptively encode and predict sensory inputs from natural scenes (e.g., [23, 37, 339]. The goal of this review is to map such recurrent connectivity rules with the computational blocks and their dynamics. Thus, in Sec. 4 (see also Tables 2 and 1), we will recap some key papers from the biological vision literature in a task centric manner in order to show how critical information gathered at different scales and different context can be used to design innovative and performing algorithms.

In the context of the long-lasting debate about the precise relationships between structures and functions, we shall briefly mention the recent attempts to derive deeper insight about the processing hierarchy along the cortical ventral pathway. It has been suggested that deep convolutional neural networks (DCNNs) provide a potential framework for modelling biological vision. A directly related question is degree of similarity between the learning process implemented over several hierarchies in order to build feature layers of different selectivities with the cellular functional properties that have been identified in different cortical areas [167]. One proposal to generate predictive models of visual cortical function along the ventral path utilises a goal-driven approach to deep learning [379]. In a nutshell, such an approach optimises network parameters regarding performance on a task that is behaviourally relevant and then compares the resulting network(s) against neural data. As emphasised here, a key element in such a structural learning approach is to define the task-level properly and then map principled operations of the system onto the structure of the system. In addition, several parameters of deep networks are usually

defined by hand, such a the number of layers or the number of feature maps within a layer. There have been recent proposals to optimise these automatically, e.g., by extensive searching or using genetic algorithms [262, 28].

3.5 Testing biologically-inspired models against both natural and computer vision

The dynamics of the biological visual systems have been probed at many different levels, from the psychophysical estimation of perceptual or behavioural performance to the physiological examination of neuronal and circuits properties. This diversity has led to a fragmentation of computational models, each targeting a specific set of experimental conditions, stimuli or responses.

Let consider visual motion processing in order to illustrate our point. When both neurally and psychophysically motivated models have been developed for a specific task such as motion integration for instance, they have been tested using a limited set of non-naturalistic inputs such as moving bars, gratings and plaid patterns (e.g., [240, 301]). These models formalise empirical laws that can explain either the perceived direction or the emergence of neuronal global motion direction preference. However, these models are hardly translated to velocity estimations in naturalistic motion stimuli since they do not handle scenarios such as lack of reliable cues or extended motion boundaries. By consequence, these models are very specific and not applicable directly to process generic motion stimuli. To overcome this limitation, a few extended computational models have been proposed that can cope with a broader range of inputs. These computational models handle a variety of complex motion inputs [117, 340] but the specific algorithms have been tuned to recover coarse attributes of global motion estimation such as the overall perceived direction or the population neuronal dynamics. Such tuning strongly limits their ability to solve tasks such as dense optical flow estimation. Still, their computational principles can be used as building blocks to develop extended algorithms that can handle naturalistic inputs [253, 326]. Moreover, they can be evaluated against standard computer vision benchmarks [15, 51]. What is still missing are detailed physiological and psychophysical data collected with complex scenarios such as natural or naturalistic images in order to be able to further constrain these models.

A lesson to be taken from the above example is that a successful synergistic approach between artificial and natural vision should first establish a common set of naturalistic inputs against which both bio-inspired and computer vision models can be benchmarked and compared. This step is indeed critical for identifying scenarios in which biological vision systems deviate with respect to the definition adopted by the computer vision. On the other side, state-of-the-art computer vision algorithms shall also be evaluated relative to human perception performance for the class of stimuli widely used in psychophysics. For the three illustrative tasks to be discussed below, we will show the interest of common benchmarks for comparing biological and computer vision solutions.

3.6 Task-based versus general purpose vision systems

Several objections can be raised to question the need for a synergy between natural and biological vision. A first objection is that biological and artificial systems could serve different aims. In particular, the major aim of biological vision studies is to understand the behaviours and properties of a general purpose visual system that could subserve different types of perceptions or actions. This generic, encapsulated visual processing machine can then be linked with other cognitive systems in an adaptive and flexible way (see [276, 344] for example). By contrast, computer vision approaches are more focused on developing task specific solutions, with an

ever growing efficiency thank to advances in algorithms (e.g., [178, 225]) supported by growing computing power. A second objection is that the brain might not use the same general-purpose (Euclidean) description of the world that Marr postulated [368]. Thus perception may not use the same set of low-level descriptors as computer vision, dooming the search for common early algorithms. A third, more technical objection is related to the low performance of most (if not all) current bio-inspired vision algorithms when solving a specific task (e.g., face recognition) when compared to state-of-the-art computer vision solutions. Moreover, bio-inspired models are still too often based on over-simplistic inputs and conditions and not sufficiently challenged with high-dimension inputs such as complex natural scenes or movies. Finally, artificial systems can solve a particular task with a greater efficiency than human vision for instance, challenging the need for bio-inspiration.

These objections question the interest of grounding computer vision solution on biology. Still, many other researchers have argued that biology can help recasting ill-based problems and showing us to ask the right questions and identifying the right constraints [387, 346]. Moreover, to mention one recent example, perceptual studies can still identify feature configurations that cannot be used by current models of object recognition and thus reframing the theoretical problems to be solved to match human performance [353]. Finally, recent advances in computational neurosciences has identified generic computational modules that can be used to solve several different perceptual problems such as object recognition, visual motion analysis or scene segmentation, just to mention a few (e.g. [57, 68, 95]). Thus, understanding task-specialised subsystems by building and testing them remains a crucial step to unveil the computational properties of building blocks that operate in largely unconstrained scene conditions and that could later be integrated into larger systems demonstrating enhanced flexibility, default-resistance or learning capabilities. Theoretical studies have identified several mathematical frameworks for modelling and simulating these computational solutions that could be inspiring for computer vision algorithms. Lastly, current limitations of existing bio-inspired models in terms of their performance will also be solved by scaling up and tuning them such that they pass the traditional computer vision benchmarks.

We propose herein that the task level approach is still an efficient framework for this dialogue. Throughout the next sections, we will illustrate this standpoint with three particular examples: retinal image sensing, scene segmentation and optic flow computation. We will highlight some important novel constraints emerging from recent biological vision studies, how they have been modelled in computational vision and how they can lead to alternative solutions.

4 Solving vision tasks with a biological perspective

In the preceding sections, we have revisited some of the main features of biological vision and we have discussed the foundations of the current computational approaches of biological vision. A central idea is the functional importance of the task at hand when exploring or simulating the brain. Our hypothesis is that such a task centric approach would offer a natural framework to renew the synergy between biological and artificial vision. We have discussed several potential pitfalls of this task-based approach for both artificial and bio-inspired approaches. But we argue that such task-centric approach will escape the difficult, theoretical question of designing general-purpose vision systems for which no consensus is achieved so far in both biology and computer vision. Moreover, this approach allow us to benchmark the performance of computer and bio-inspired vision systems, an essential step for making progress in both fields. Thus, we believe that the task-based approach remains the most realistic and productive approach. The novel strategy based on bio-inspired generic computational blocks will however open the door for improving the scalability, the flexibility and the fault-tolerance of novel computer vision solutions. As already stated above, we decided to revisit three classical computer vision tasks from such a biological perspective: image sensing, scene segmentation and optical flow.[1] This choice was made in order to provide a balanced overview of recent biological vision studies about three illustrative stages of vision, from the sensory front-end to the ventral and dorsal cortical pathways. For these three tasks, there are a good set of multiple scales biological data and a solid set of modelling studies based on canonical neural computational modules. This enables us to compare these models with computer vision algorithms and to propose alternative strategies that could be further investigated. For the sake of clarity, each task will be discussed with the following framework:

Task definition. We start with a definition of the visual processing task of interest.

Core challenges. We summarise its physical, algorithmic or temporal constraints and how they impact the processing that should be carried on images or sequences of images.

Biological vision solution. We review biological facts about the neuronal dynamics and circuitry underlying the biological solutions for these tasks stressing the canonical computing elements being implemented in some recent computational models.

Comparison with computer vision solutions. We discuss some of the current approaches in computer vision to outline their limits and challenges. Contrasting these challenges with known mechanisms in biological vision would be to foresee which aspects are essential for computer vision and which ones are not.

Promising bio-inspired solutions. Based on this comparative analysis between computer and biological vision, we discuss recent modelling approaches in biological vision and we highlight novel ideas that we think are promising for future investigations in computer vision.

4.1 Sensing

Task definition. Sensing is the process of capturing patterns of light from the environment so that all the visual information that will be needed downstream to cater the computational/functional needs of the biological vision system could be faithfully extracted. This

[1] See also, recent review articles addressing other tasks: object recognition [7], visual attention [344, 347], biological motion [105].

definition does not necessarily mean that its goal is to construct a veridical, pixel-based representation of the environment by passively transforming the light the sensor receives.

Core challenges. From a functional point of view, the process of sensing (i.e., transducing, transforming and transmitting) light patterns encounters multiple challenges because visual environments are highly cluttered, noisy and diverse. First, illumination levels can vary over several range of magnitudes. Second, image formation onto the sensor is sensitive to different sources of noise and distortions due to the optical properties of the eye. Third, transducing photons into electronic signals is constrained by the intrinsic dynamics of the photosensitive device, being either biological or artificial. Fourth, transmitting luminance levels on a pixel basis is highly inefficient. Therefore, information must be (pre-)processed so that only the most relevant and reliable features are extracted and transmitted upstream in order to overcome the limited bandpass properties of the optic nerve. At the end of all these different stages, the sensory representation of the external world must still be both energy and computationally very efficient. All these aforementioned aspects raise some fundamental questions that are highly relevant for both modelling biological vision and improving artificial systems.

Herein, we will focus on four main computational problems (what is computed) that are illustrative about how biological solutions can inspire a better design of computer vision algorithms. The first problem is called *adaptation* and explains how retinal processing is adapted to the huge local and global variations in luminance levels from natural images in order to maintain high visual sensitivity. The second problem is *feature extraction*. Retinal processing extracts information about the structure of the image rather than mere pixels. What are the most important features that sensors should extract and how they are extracted are pivotal questions that must be solved to sub-serve an optimal processing in downstream networks. Third is the *sparseness* of information coding. Since the amount of information that can be transmitted from the front-end sensor (the retina) to the central processing unit (area V1) is very limited, a key question is to understand how spatial and temporal information can be optimally encoded, using context dependency and predictive coding. The last selected problem is called *precision* of the coding, in particular what is the temporal precision of the transmitted signals that would best represent the seaming-less sequence of images.

Biological vision solution. The retina is one of the most developed sensing devices [110, 206, 207]. It transforms the incoming light into a set of electrical impulses, called spikes, which are sent asynchronously to higher level structures through the optic nerve. In mammals, it is sub-divided into five layers of cells (namely, photoreceptors, horizontal, bipolar, amacrine and ganglion cells) that forms a complex recurrent neural network with feedforward (from photoreceptors to ganglion cells), but also lateral (i.e., within bipolar and ganglion cells layers) and feedback connections. The complete connectomics of some invertebrate and vertebrate retinas now begin to be available [199].

Regarding information processing, an humongous amount of studies have shown that the mammalian retina can tackle the four challenges introduced above using *adaptation*, *feature detection*, *sparse coding* and *temporal precision* [156]. Note that *feature detection* should be understood as "feature encoding" in the sense that there is non decision making involved. Concerning *adaptation*, it is a crucial step, since retinas must maintain high contrast sensitivity over a very broad range of luminance, from starlight to direct sunlight. Adaptation is both global through neuromodulatory feedback loops and local through adaptive gain control mechanisms so that retinal networks can be adapted to the whole scene illuminance level while maintaining high contrast sensitivity in different regions of the image, despite their considerable differences in luminance [314, 75, 336].

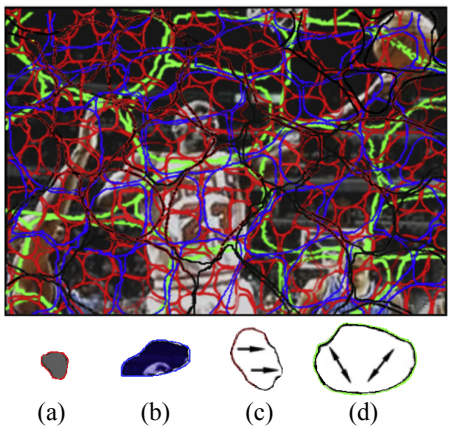


Figure 4: How retinal ganglion cells tile a scene extracting a variety of features. This illustrates the tiling of space of a subset of four cell types. Each tile covers completely the visual image independently from other types. The four cell types shown here correspond to (a) cell with small receptive fields and center-surround characteristics extracting intensity contrasts, (b) color coded cells, (c) motion direction selective cells with a relatively large receptive field, (d) cells with large receptive fields reporting that something is moving (adapted from [207], with permissions).

It has long been known that retinal ganglion cells extract local luminance profiles. However, we have now a more complex view of retinal form processing. The retina of higher mammals sample each point in the images with about 20 distinct ganglion cells [206, 207] associated to different *features*. This is best illustrated in Fig. 4, showing how the retina can gather information about the structure of the visual scene with four example cell types tilling the image. They differ one from the others by the size of their receptive field and their spatial and temporal selectivities. These spatiotemporal differences are related to the different sub-populations of ganglion cells which have been identified. Parvocellular (P) cells are the most numerous are the P-cells (80%). They have a small receptive size and a slow response time resulting in a high spatial resolution and a low temporal sensitivity. They process information about color and details. Magnocellular cells have a large receptive field and a low response time resulting in a high temporal resolution and a low spatial sensitivity, and can therefore convey information about visual motion [313]. Thus visual information is split into parallel stream extracting different domains of the image spatiotemporal frequency space. This was taken at a first evidence for feature extractions at retinal level. More recent studies have shown that, in many species, retinal networks are much smarter than originally thought. In particular, they can extract more complex features such as basic static or moving shapes and can predict incoming events, or adapt to temporal changes of events, thus exhibiting some of the major signatures of predictive coding [110, 206, 207].

A striking aspect of retinal output is its *high temporal precision* and *sparseness*. Massive in vitro recordings provide spiking patterns collected from large neuronal assemblies so that it becomes possible to decipher the retinal encoding of complex images [260]. Modelling the spiking output of the ganglion cell populations have shown high temporal precision of the spike trains and a strong reliability across trials. These coding properties are essential for upstream processing what will extract higher order features but also will have to maintain such high precision. In brief, the retina appears to be a dense neural network where specific sub-populations adaptively extract local information in a context-dependent manner in order to produce an output that is both adaptive, sparse, over complete and of high temporal precision.

Another aspect of retinal coding is its space-varying resolution. A high-resolution sampling zone appears in the fovea while the periphery looses spatial detail. The retinotopic mapping of receptors into the cortical representation can be characterized formally by a non-linear conformal mapping operation. Different closed-form models have been proposed which share the property that the retinal image is sampled in a space-variant fashion using a topological transformation of the retinal image into the cortex. The smooth variation of central into peripheral vision may directly support a mechanism of space-variant vision. Such active processing mechanism not only significantly reduces the amount of data (particularly with a high rate of peripheral compression) but may also support computational mechanisms, such as symmetry and motion detection.

There is a large, and expanding body of literature proposing models of retinal processing. We attempted to classify them and isolated three main classes of models. The first class regroups the linear-nonlinear-poisson (LNP) models [241]. In its simplest form, a LNP model is a convolution with a spatio-temporal kernel followed by a static nonlinearity and stochastic (Poisson-like) mechanisms of spikes generation. These functional model are widely used by experimentalists to characterise the cells that they record, map their receptive field and characterise their spatiotemporal feature selectivities [62]. LNP models can simulate the spiking activity of ganglion cells (and of cortical cells) in response to synthetic or natural images [56] but they voluntarily ignore the neuronal mechanisms and the details of the inner retinal layers that transform the image into a continuous input to the ganglion cell (or any type of cell) stages. Moreover, they implement static non-linearities, ignoring many existing non-linearities. Applied to computer vision, they however provide some inspiring computational blocks for contrast enhancement, edge detection or texture filtering.

The second class of models has been developed to serve as a front-end for subsequent computer vision task. They provide bio-inspired modules for low level image processing. One interesting example is given by [26, 129], where the model includes parvocellular and magnocellular pathways using different non-separable spatio-temporal filter that are optimal for form or motion detection.

The third class is based on detailed retinal models reproducing its circuitry, in order to predict the individual or collective responses measured at the ganglion cells level [375, 192]. Virtual Retina [375] is one example of such spiking retina model. This models enables large scale simulations (up to 100,000 neurons) in reasonable processing times while keeping a strong biological plausibility. These models are expanded to explore several aspects of retinal image processing such as (i) understanding how to reproduce accurately the statistics of the spiking activity at the population level [233], (ii) reconciling connectomics and simple computational rules for visual motion detection [160] and (iii) investigating how such canonical microcircuits can implement the different retinal processing modules cited above (feature extraction, predictive coding) [110].

Comparison with computer vision solutions. Most computer vision systems are rooted on a sensing device based on CMOS technology to acquire images in a frame based manner. Each frame is obtained from sensors representing the environment as a set of pixels whose values indicate the intensity of light. Pixels pave homogeneously the image domain and their number defines the resolution of images. Dynamical inputs, corresponding to videos are represented as a set of frames, each one representing the environment at a different time, sampled at a constant time step defining the frame rate.

To make an analogy between the retina and typical image sensors, the dense pixels which respond slowly and capture high resolution color images are at best comparable to P-Cells in the retina. Traditionally in computer vision, the major technological breakthroughs for sensing devices have aimed at improving the density of the pixels, as best illustrated by the ever improving resolution of the images we capture daily with cameras. Focusing of how videos are captured,

one can see that a dynamical input is not more that a series of images sampled at regular intervals. Significant progress have been achieved recently in improving the temporal resolution with advent of computational photography but at a very high computational cost [187]. This kind of sensing for videos introduces a lot of limitations and the amount of data that has to be managed is high.

However, there are two main differences between the retina and a typical image sensor such as a camera. First, as stated above, the retina is not simply sending an intensity information but it is already extracting features from the scene. Second, the retina asynchronously processes the incoming information, transforming it as a continuous succession of spikes at the level of ganglion cells, which mostly encode changes in the environment: retina is very active when intensity is changing, but its activity becomes quickly very low with a purely static stimulation. These observations show that the notion of representing static frames does not exist in biological vision, drastically reducing the amount of data that is required to represent temporally varying content.

Promising bio-inspired solutions. Analysing the sensing task from a biological perspective has potential for bringing new insights and solutions related to the four challenges outlined in this section. In terms of an ideal sensor, it is desired to have control over the acquisition of each pixel, thus allowing a robust adaptation to different parts of the scene. However, this is difficult to realize on the chip as it would mean independent triggers to each pixel, thus increasing the information transfer requirements on the sensor. In order to circumvent this problem, current CMOS sensors utilize a global clock trigger which fails us to give a handle on local adaptation, thus forcing a global strategy. This problem is tackled differently in biologically inspired sensors, by having local control loops in the form of event driven triggering rather than a global clock based drive. This helps the sensor to adapt better to local changes and avoids the need for external control signals. Also, since the acquisitions are to be rendered, sensory physiological knowledge could help in choosing good tradeoffs on sensor design. For example, the popular Bayer filter pattern has already been inspired by the physiological properties of retinal color sensing cells. With the advent of high dynamic range imaging devices, these properties are beginning to find interesting applications such as low range displays. This refers to the tone mapping problem. It is a necessary step to visualize high-dynamic range images on low-dynamic range displays, spanning up to two orders of magnitude. There is a large body of literature in this area on static images (see [171, 29] for reviews), with approaches which combine luminance adaptation and local contrast enhancement sometimes closely inspired from retinal principles, as in [219, 25, 90, 227] just to cite a few. Recent developments concern video-tone mapping where a few approaches have been developed so far (see [83] for a review). We think it is for videos that the development of synergistic models of the retina is the most promising. Building on existing detailed retinal models such as the Virtual Retina [375] (mixing filter-based processing, dynamical systems and spiking neuron models), the goal is to achieve a better characterization of retinal response dynamics which will have a direct application here.

The way that retina performs *feature detection* and encodes information in space and time has received relatively little attention so far from the computer vision community. In most cases, retina-based models rely on simple caricatures of the retina. The FREAK (Fast Retina Keypoint) descriptor [4] is one example where only the geometry and space-varying resolution has been exploited. In [4], the "cells" in the model are only doing some averaging of intensities inside their receptive field. This descriptor model was extended in [132] where ON and OFF cells were introduced using a linear-nonlinear (LN) model. This gives a slight gain of performance in a classification task, although it is still far from the state-of-the-art. These descriptors could be improved in many ways, by taking into account the goal of the features detected by the 20

types of ganglion cells mentioned before. Here also the strategy is to build on existing retinal models. In this context, one can also mention the SIFT descriptor [193] which was also inspired by cortical computations. One needs to evaluate the functional implication at a task level of some retinal properties. Examples include the asymmetry between ON and OFF cells [248] and the irregular receptive field shapes [190].

One question is whether we would still need inspiration from the retina to build new descriptors, given the power of machine learning methods that provides automatically some optimized features given an image database? What the FREAK-based models show is that it is not only about improving the filters. It is also about how the information is encoded. In particular, what is encoded in FREAK-based models is the relative difference between cell responses. Interestingly, this is exactly the same as the rank-order coding idea proposed as an efficient strategy to perform ultra-fast categorization [359], and which has been reported in the retina [268]. This idea has been exploited for pattern recognition and used in many applications as demonstrated by the products developed by the company Spikenet (http://www.spikenet-technology.com). This means that the retina should serve as a source of inspiration not only to propose features, but more importantly, how it encodes these features at a population level.

The fact that the retinal output is *sparse* and has a *high temporal precision* conveys a major advantage to the visual system, since it has to deal with only a small amount of information. A promising bio-inspired solution is to develop frame-free methods, i.e., methods using sparse encoding of the visual information. This is now possible using event-based vision sensors where pixels autonomously communicate the change and grayscale events. The dynamic vision sensor (DVS) [184, 188] and the asynchronous time-based image sensor (ATIS) [269] are two examples of such sensor using address-event representation (AER) circuits. The main principle is that pixels signal only significant events. More precisely, an event is sent when the log intensity has changed by some threshold amount since the last event (see Fig. 5). These sensors provide a sparse output corresponding to pixels that register a change in the scene, thus allowing extremely high temporal resolution to describe changes in the scene while discarding all the redundant information. Because the encoding is sparse, these sensors appear as a natural solution in real-time scenarios or when energy consumption is a constraint. Combined with what is known about retinal circuitry as in [192], they could provide a very efficient front-end for subsequent visual tasks, in the same spirit of former neuromorphic models of low-level processing as in [26, 129]. They could also be used more directly as a way to represent visual scenes, abandoning the whole notion of a video that is composed of frame-sequences. This provides a new operative solution that can be used to revisit computer vision problems (see [189] for a review). This field is rapidly emerging, with the motivation to develop approaches more efficient than the state-of-the-art. Some examples include tracking [236], stereo [296], 3D pose estimation [357], object recognition [245] and optical flow [27, 342, 45, 109].

4.2 Segmentation and figure-ground segregation

Task definition. The task of segmenting a visual scene is to generate a meaningful partitioning of the input feature representation into surface- or object-related components. The segregation of an input stimulus into prototypical parts, characteristic of surfaces or objects, is guided by a coherence or homogeneity property that region elements share. Homogeneities are defined upon feature domains such as color, motion, depth, statistics of luminance items (texture), or combinations of them [247, 204]. The specificity of the behavioural task, e.g., grasping an object, distinguishing two object identities, or avoiding collisions during navigation, may influence the required detail of segmentation [16, 122]. In order to do so, contextual information in terms of high-level knowledge representations can be exploited as well [32]. In addition, the goal of

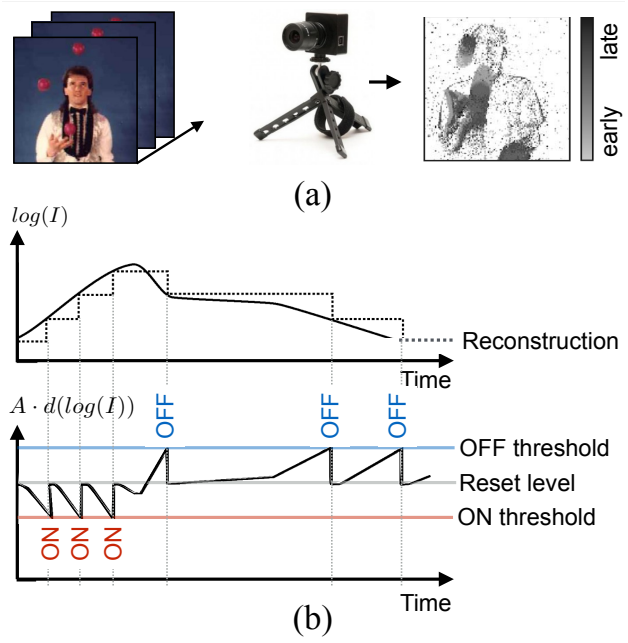


Figure 5: How DVS sensor generate spikes. (a) Example of a video with fast motions (a juggling scene). DVS camera and DVS output: Events are rendered using a grayscale colormap corresponding to events that were integrated over a brief time window (black = young, gray = old, white = no events). (b) DVS principle: Positive and negative changes are generated depending on the variations of $log(I)$ which are indicated as ON and OFF events along temporal axis (adapted from [184], with permissions).

segmentation might be extended in regard to eventually single out a target item, or object, from its background in order to recognise it or to track its motion.

Core challenges. The segmentation of a spatio-temporal visual image into regions that correspond to prototypical surfaces or objects faces several challenges which derive from distinct interrelated subject matters. The following themes refer to issues of *representation*. First, the feature domain or multiple domains need to be identified which constitute the coherence or homogeneity properties relevant for the segregation task. Feature combinations as well as the nested structure of their appearance of coherent surfaces or objects introduces apparent feature hierarchies [161, 162]. Second, the segmentation process might focus on the analysis of homogeneities that constitute the coherent components within a region or, alternatively, on the discontinuities between regions of homogeneous appearances. Approaches belonging to the first group focus on the segregation of parts into meaningful prototypical regions utilising an agglomeration (clustering) principle. Approaches belonging to the second group focus on the detection of discontinuous changes in feature space (along different dimensions) [239] and group them into contours and boundaries. Note that we make a distinction here to refer to a contour as a grouping of oriented edge or line contrast elements whereas a boundary already relates to a surface border in the scene. Regarding the boundaries of any segment, the segmentation task itself might incorporate an explicit assignment of a border ownership (BOwn) direction label which implies the separation of figural shape from background by a surface that occludes other scenic parts [256, 164]. The

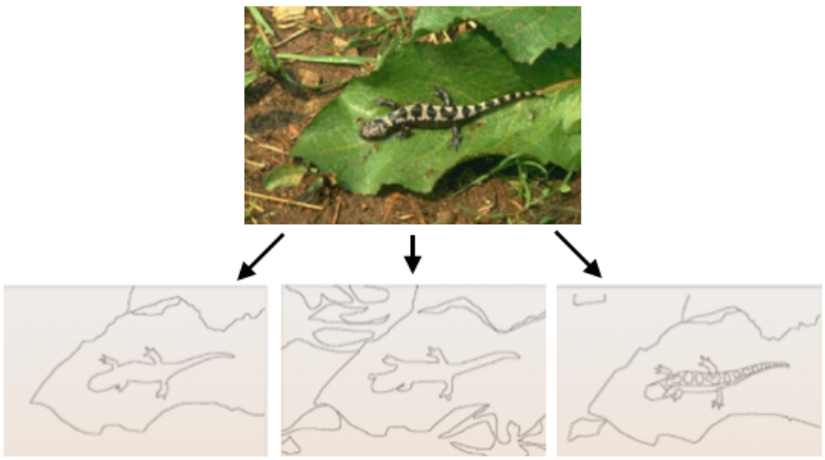


Figure 6: Example of possible segmentation results for a static image drawn by different human observers. Lower images shows segmentations happening at different levels of detail but consistent with each other (adapted from [11]).

variabilities in the image acquisition process caused by, e.g., illumination conditions, shape and texture distortions, might speak in favor of a boundary oriented process. On the other hand, the complexity of the background structure increases the effort to segregate a target object from the background, which argues in favour of region oriented mechanisms. It should be noted, however, that the region vs boundary distinction might not appear as binary as in the way outlined above. Considering real world scenes the space-time relationships of perceptual elements (defined over different levels of resolution) are often defined by statistically meaningful structural relations to determine segmentation homogeneities [374]. Here, an important distinction has been made between structure that might be influenced by meaning and primitive structure that is perceived even without a particular interpretation.

While the previous challenges were defined by representations, the following themes refer to the process characteristic of segmentation. First, the partitioning process may yield different results given changing view-points or different noise sources during the sensing process. Thus, segmentation imposes an inference problem that is mathematically ill-posed [265]. The challenge is how a reliability, or confidence, measure is defined that characterises meaningful decompositions relating to reasonable interpretations. To illustrate this, Fig. 6 shows segmentation results as drawn by different human observers. Second, figural configurations may impose different efforts for mechanisms of perceptual organisation to decide upon the segregation of an object from the background and/or the assignment of figure and ground direction of surface boundaries. A time dependence that correlates with the structural complexity of the background has in fact been observed to influence the temporal course needed in visual search tasks [376].

Biological vision solution. Evidence from neuroscience suggests that the visual system uses segmentation strategies based on identifying discontinuities and grouping them into contours and boundaries. Such processes operate mainly in a feedforward fashion and automatic, utilising early and intermediate-level stages in visual cortex. In a nutshell, contrast and contour detection is quickly accomplished and is already represented at early stages in the visual cortical hierarchy, namely areas V1 and V2. The assignment of task-relevant segments happens to occur after a slight temporal delay and involves a recurrent flow of lateral and feedback processes [291, 306,

292].

The grouping of visual elements into contours appears to follow the Gestalt rules of perceptual organisation [163]. Grouping has also been studied in accordance to the ecological validity of such rules as they appear to be embedded in the statistics of natural scenes [47]. Mechanisms that entail contour groupings are implemented in the structure of supragranular horizontal connections in area V1 in which oriented cells preferentially contact like-oriented cells that are located along the orientation axes defined by a selected target neuron [154, 36]. Such long-range connections form the basis for the Gestalt concept of good continuation and might reflect the physiological substrate of the association field, a figure-eight shaped zone of facilitatory coupling of orientation selective input and perceptual integration into contour segments [114, 91, 101]. Recent evidence suggests that the perceptual performance of visual contour grouping can be improved by mechanisms of perceptual learning [182]. Once contours have been formed they need to be labelled in accordance to their scene properties. In case of a surface partially occluding more distant scenic parts the *border ownership (BOwn) or surface belongingness* can be assigned to the boundary [163]. A neural correlate of such a mechanism has been identified at different cortical stages along the ventral pathway, such as V1, V2 and V4 areas [385, 242]. The dynamics of the generation of the BOwn signals may be explained by feedforward, recurrent lateral and feedback mechanisms (see [373] for a review).

Such dynamical process of feedback, called re-entry [80], recursively links representations distributed over different levels. Mechanisms of lateral integration, although slower in processing speed, seem to further support intra-cortical grouping [154, 155, 108]. In addition, surface segregation is reflected in a later temporal processing phase but is also evident in low levels of the cortical hierarchy, suggesting that recurrent processing between different cortical stages is involved in generating neural surface representations. Once boundary groupings are established surface-related mechanisms "paint", or tag, task-relevant elements within bounded regions. The feature dimensions used in such grouping operations are, e.g., local contour orientations defined by luminance contrasts, direction and speed of motion, color hue contrasts, or texture orientation gradients. As sketched above, counter-stream interactive signal flow [351] imposes a temporal signature on responses in which after a delay a late amplification signal serves to tag those local responses that belong to a region (surrounded by contrasts) which has been selected as a figure [172] (see also [295]). The time course of the neuronal responses encoding invariance against different figural sizes argues for a dominant role of feedback signals when dynamically establishing the proper BOwn assignment. Grouping cells have been postulated that integrate (undirected) boundary signals over a given radius and enhance those configurations that define locally convex shape fragments. Such fragments are in turn enhanced via a recurrent feedback cycle so that closed shape representations can be established rapidly through the convexity in closed bounding contours [385]. Neural representations of localized features composed of multiple orientations may further influence this integration process, although this is not firmly established yet [10]. BOwn assignment serves as a prerequisite of figure-ground segregation. The temporal dynamics of cell responses at early cortical stages suggest that mechanisms exist that (i) decide about ownership direction and (ii) subsequently enhance regions (at the interior of the outline boundaries) by spreading a neural tagging, or labelling, signal that is initiated by the region boundary [294] (compare the discussion in [373]). Such a late enhancement through response modulation of region components occurs for different features, such as oriented texture [173] or motion signals [295], and is mediated by recurrent processes of feedback from higher levels in the cortical hierarchy. It is, however, not clear whether a spreading process for region tagging is a basis for generating invariant neural surface representations in all cases. All experimental investigations have been conducted for input that leads to significant initial stimulus responses while structure-less homogeneous regions (e.g., a homogeneous coloured wall) may lead to void

spaces in the neuronal representation that may not be filled explicitly by the cortical processing (compare the discussion in [254]).

Yet another level of visual segmentation operates upon the initial grouping representations, those base groupings that happen to be processed effortlessly as outlined above. However, the analysis of complex relationships surpasses the capacities of the human visual processor which necessitates serial staging of some higher-level grouping and segmentation mechanisms to form incremental task-related groupings. In this mainly sequential operational mode visual routines establish properties and relations of particular scene items [350]. Elemental operations underlying such routines have been suggested, e.g., shifting the processing focus (related to attentional selection), indexing (to select a target location), coloring (to label homogeneous region elements), and boundary tracing (determining whether a contour is open or closed and items belonging to a continuous contour). For example, contour tracing is suggested to be realized by incremental grouping operations which propagate an enhancement of neural firing rates along the extent of the contour. Such a neural labelling signal is reflected in a late amplification in the temporal signature of neuronal responses. The amplification is delayed with respect to the stimulus onset time with increasing distances of the location along the perceptual entity [151, 292] (that is indexed by the fixation point at the end of the contour). This lead to the conclusion that such tracing is laterally propagated (via lateral or interative feedforward and feedback mechanisms), leading to a neural segmentation of the labelled items delineating feature items that belong to the same object or perceptual unit. Maintenance operations then interface such elemental operations into sequences to compose visual routines for solving more complex tasks, like in a sequential computer program. Such cognitive operations are implemented in cortex by networks of neurons that span several cortical areas [290]. The execution time of visual cortical routines reflects the sequential composition of such task-specific elemental neural operations tracing the signature of neural responses to a stimulus [175, 290].

Comparison with computer vision solutions. Segmentation as an intermediate level process in computational vision is often characterised as one of agglomerating, or clustering, picture elements to arrive at an abstract description of the regions in a scene [247]. It can also be viewed as a preprocessing step for object detection/recognition. It is not very surprising to see that even in computer vision earlier attempts were drawn towards single aspects of the segmentation like edge detection [203, 54, 185] or grouping homogeneous regions by clustering [65]. The performance limitations of both these approaches independently have led to the emergence of solutions that reconsidered at the problem as a juxtaposition of both edge detection and homogeneous region grouping with implicit consideration for scale. The review paper by [96] presents various approaches that attempted in merging edge based information and clustering based information in a sequential or parallel manner. The state of the art techniques that are successful in formulating the combined approach are variants of graph cuts [317], active contours, and level sets. At the bottom of all such approaches is the definition of an optimisation scheme that seeks to find a solution under constraints such as, e.g., smoothness or minimising a measure of total energy. These approaches are much better in terms of meeting human defined ground truth compared to simpler variants involving discontinuity detection or clustering alone. The performance of computer vision approaches to image partitioning has been boosted recently by numerous contributions utilizing DCNNs for segmentation (e.g., [238, 138, 139]). The basic structure of the encoder component of segmentation networks is similar to the hierarchical networks trained for object recognition [168]. For example, the *AlexNet* has been trained by learning a hierarchy of kernels in the convolutional layers to extract rich feature sets for recognition from a large database of object classes. Segmentation networks [238, 138] have been designed by adding a decoder scheme to expand the activations in the category layers through a sequence of decon-

volutions steps such as in autoencoder networks [134]. Even more extended versions include a mechanism of focused attention to more selectively guide the training process using class labels or segmentations [139]. The hierarchical structure of such approaches shares several features of cortical processing through a sequence of areas with cells that increase their response selectivity at the size of their receptive fields over different stages in the cortical hierarchy. However, the explicit unfolding of the data representation in the deconvolution step to upscale to full image resolution, the specific indexing of pixel locations to invert the pooling in the deconvolution, and the large amount of training data are not biologically plausible.

A major challenge is still how to compare the validity and the quality of segmentation approaches. Recent attempts emphasise to compare the computational results - from operations on different scales - with the results of hand-drawn segmentations by human subjects [94, 11]. These approaches suggest possible measures in judging the quality of automatic segmentation given that ground truth data is missing. However, the human segmentation data does not elucidate the mechanisms underlying the processes to arrive at such partitions. Instead of a global partitioning of the visual scene, the visual system seems to adopt different strategies of computation to arrive at a meaningful segmentation of figural items. The grouping of elements into coherent form is instantiated by selectively enhancing the activity of neurons that represent the target region via a modulatory input from higher cortical stages [172, 174]. The notion of feedback to contribute in the segmentation of visual scenes has been elucidated above. Recent computer vision algorithms begin to make use of such recurrent mechanisms as well. For example, since bottom-up data-driven segmentation is usually incomplete and ambiguous the use of higher-level representations might help to validate initial instances and further stabilise their representation [352, 32]. Along this line, top-down signalling applies previously acquired information about object shape (e.g., through learning), making use of the discriminative power of fragments of intermediate size, and combines this information with a hierarchy of initial segments [354]. Combined contour and region processing mechanisms have also been suggested to guide the segmentation. In [11], multi-scale boundaries are extracted which later prune the contours in a watershed region-filling algorithm. Algorithms of figure-ground segregation and border-ownership computation have been developed for computer vision applications to operate under realistic imaging conditions [328, 332]. These were designed to solve tasks like shape detection against structured background and for video editing. Still, the robust segmentation of an image into corresponding surface patches is hard to accomplish in a reliable fashion. Performance of such methods mentioned above depends on parametrization and the unknown complexity and properties of the viewed scene. Aloimonos and coworkers proposed an active vision approach that adopted biological principles like the selection and fixation on image regions that are surrounded by closed contours [223, 224]. The key here is that in this approach only the fixated region (corresponding to a surface of an object or the object itself) is then segmented based on an optimization scheme using graph-cut. All image content outside the closed region contour is background w.r.t. the selected target region or object. The functionality requires an active component to relocate the gaze and a region that is surrounded by a contrast criterion in the image.

Promising bio-inspired solutions. Numerous models that account for mechanisms of contour grouping have been proposed to linking orientation selective cells [114, 116, 183]. The rules of mutual support utilize a similarity metric in the space-orientation domain giving rise to a compatibility, or reliability measure [157] (see [235] for a review of generic principles and a taxonomy). Such principles migrated into computer vision approaches [249, 216, 165] and, in turn, provided new challenges for experimental investigations [318, 24]. Note that the investigation of structural connectivities in high dimensional feature spaces and their mapping onto a low-

dimensional manifold lead to define a "neurogeometry" and the basic underlying mathematical principles of such structural principles [257, 63].

As outlined above, figure-ground segregation in biological vision segments an image or temporal sequence by boundary detection and integration followed by assigning border ownership direction and then tagging the figural component in the interior of a circumscribed region. Evidence suggests that region segmentation by tagging the items which belong to extended regions involves feedback processing from higher stages in the cortical hierarchy [306]. Grossberg and colleagues proposed the FACADE theory (form-and-color-and-depth [114, 113]) to account for a large body of experimental data, including figure-ground segregation and 3D surface perception. In a nutshell, the model architecture consists of mutually coupled subsystems, each one operating in a complementary fashion. A boundary contour system (BCS) for edge grouping is complemented by a feature contour system (FCS) which supplements edge grouping by allowing feature qualities, such as brightness, color, or depth, to spread within bounded compartments generated by the BCS.

The latter mechanism has recently been challenged by psychophysical experiments that measure subject reaction times in image-parsing tasks. The results suggest that a sequential mechanism groups, or tags, interior patches along a connected path between the fixation spot and a target probe. The speed of reaching a decision argues in favor of a spreading growth-cone mechanism that simultaneously operates over multiple spatial scales rather than the wave-like spreading of feature activities initiated from the perceptual object boundary [149]. Such a mechanism is proposed to also facilitate the assignment of figural sides to boundaries. BOwn computation has been incorporated in computer vision algorithms to segregate figure and background regions in natural images or scenes [284, 137, 332]. Such approaches use local configurations of familiar shapes and integrate these via global probabilistic models to enforce consistency of contour and junction configurations [284] of learning of templates from ensembles of image cues to depth and occlusion [137].

Feedback mechanisms as they are discussed above, allow to build robust boundary representations such that junctions may be reinterpreted based on more global context information [371]. The hierarchical processing of shape from curvature information in contour configurations [289] can be combined with evidence for semi-global convex fragments or global convex configurations [69]. Such activity is fed back to earlier stages of representation to propagate contextual evidences and quickly build robust object representations separated from the background. A first step towards combining such stage-wise processing capacities and integrating them with feedback that modulates activities in distributed representations at earlier stages of processing has been suggested in [341]. The step towards processing complex scenes from unconstrained camera images, however, still needs to be further investigated.

Taken together, biological vision seems to flexibly process the input in order to extract the most informative information from the optic array. The information is selected by an attention mechanism that guides the gaze to the relevant parts of the scene. It has been known for a long time that the guidance of eye movements is influenced by the observer's task of scanning pictures of natural scene content [382]. More recent evidence suggests that the saccadic landing locations are guided by contraints to optimize the detection of relevant visual information from the optic array [122, 17]. Such variability in fixation location has immediate consequences on the structure of the visual mapping into an observer representation. Consequently, segmentation might be considered as a separation problem that operates upon a high-dimensional feature space, instead of statically separating appearances into different clusters. For example, in order to separate a target object against the background in an identification task fixation is best located approximately in the middle of the central surface region [122]. Symmetric arrangement of bounding contours (with opposite direction of BOwn) helps to select the region against the

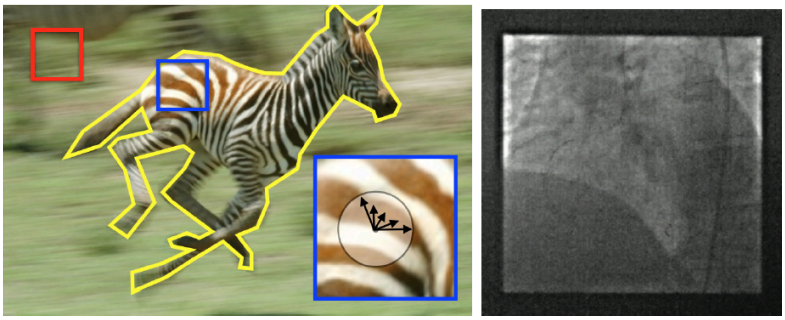
background to guide a motor action. In order to generate stable visual percept of a complex object such information must be integrated over multiple fixations [123]. In case of irregular shapes, the assignment of object belongingness requires a decision whether region elements belong to the same surface or not. Such decision-making process involves a slower sequentially operating mechanism of tracing a connecting path in a homogeneous region. Such a growth-cone mechanism has been demonstrated to act similarly on perceptual representations of contour and region representations which might tag visual elements to build a temporal signature for representations that define a connected object (compare [149]). In a different behavioral task, e.g., obstacle avoidance, the fixation close to the occluding object boundary helps to separate the optic flow pattern of the obstacle from those of the background [282]. Here, the obstacle is automatically selected as perceptual figure while the remaining visual scene structure and other objects more distant from the observer are treated as background. These examples demonstrate evidence that biological segmentation might be different from computer vision approaches which incorporates active selection elements building upon much more flexible and dynamic processes.

4.3 Optical flow

Task definition. Estimating optical flow refers to the assignment of 2-D velocity vectors at sample locations in the visual image in order to describe their displacements within the sensor's frame of reference. Such a displacement vector field constitutes the image flow representing apparent 2-D motions from their 3-D velocities being projected onto the sensor [360, 361]. These algorithms use the change of structured light in the retinal or camera images, posing that such 2-D motions are observable from light intensity variations (and thus, are contrast dependent) due to the change in relative positions between an observer (eye or camera) and the surfaces or objects in a visual scene.

Core challenges. Achieving a robust estimation of optical flow faces several challenges. First of all, visual system has to establish form-based correspondences across temporal domain despite the fact that physical movements induced geometric and photometric distortions. Second, velocity space has to be optimally sampled and represented to achieve robust and energy efficient estimation. Third, the accuracy and reliability of the velocity estimation is dependent upon the local structure/form but the visual system must achieve a form independent velocity estimation. Difficulties arise from the fact that any local motion computation faces different sources of noise and ambiguities, such as for instance the aperture and problems. Therefore, estimating optical flow requires to resolve these local ambiguities by integrating different local motion signals while still maintaining segregated those that belong to different surfaces or objects of the visual scene (see Fig. 7(a)). In other words, image motion computation faces two opposite goals when computing the global object motion, integration and segmentation [38]. As already emphasised in Sec. 4.2, any computational machinery should be able to keep segregated the different surface/object motions since one goal of motion processing is to estimate accurately the speed and direction of each of them in order to track, capture or avoid one or several of them. Fourth, the visual system must deal with complex scenes that are full of occlusions, transparencies or non-rigid motions. This is well illustrated by the transparency case. Since optical flow is a projection of 3D displacements in the world, some situations yield to perceptual (semi-) transparency [214]. In videos, several causes have been identified, such as reflections, phantom special effects, dissolve effects for a gradual shot change and medical imaging such as X-rays (for example see Fig. 7(b)). All of these examples raise serious problems to current computer vision algorithms.

Herein, we will focus on four main computational strategies used by biological systems for dealing with the aforementioned problems. We selected them because we believe these solutions



(a) (b)

Figure 7: Core challenges in motion estimation. (a) This snapshot of a moving scenes illustrates several ideas discussed in the text: inset with the blue box shows the local ambiguity of motion estimation while the yellow boundary shows how segmentation and motion estimation are intricated. (b) One example of transparent motion encountered by computer vision, from an X-ray image (from [13]).

could inspire the design of better computer vision algorithms. First is *motion energy estimation* by which the visual system estimates a contrast dependent measure of translations in order to indirectly establish correspondences. Second is *local velocity estimation*: contrast dependent motion energy features must be combined to achieve a contrast invariant local velocity estimation after de-noising the dynamical inputs and resolving local ambiguities, thanks to the integration of local form and motion cues. The third challenge concerns the *global motion estimation* of each independent object, regardless its shape or appearance. Fourth, *distributed multiplexed representations* must be used by both natural and artificial systems to segment cluttered scenes, handle multiple/transparent surfaces, and encode depth ordering to achieve 3D motion perception and goal-oriented decoding.

Biological vision solution. Visual motion has been investigated in a wide range of species, from invertebrates to primates. Several computational principles have been identified as being highly conserved by evolution, as for instance local motion detectors [121]. Following the seminal work of Werner Reichardt and colleagues, a huge amount of work has been achieved to elucidate the cellular mechanisms underlying local motion detection, the connectivity rules enabling optic flow detectors or basic figure-ground segmentation. Fly vision has been leading the investigation of natural image coding as well as active vision sensing. Several recent reviews can be found elsewhere (e.g. [34, 35, 5, 319]). In the present review, we decided to restrain the focus on the primate visual system and its dynamics. In Fig. 3, we have sketched the backbone of the primate cortical motion stream and its recurrent interactions with both area V1 and the 'form' stream. This figure illustrates both advantages and limits of the deep hierarchical model. Below, we will further focus on some recent data about the neuronal dynamics in regards with the four challenges identified for a better optic flow processing.

As already illustrated, the classical view of the cortical motion pathway is a feedforward cascade of cortical areas spanning from the occipital (V1) to the parietal (e.g. area VIP, area 7) lobes. This cascade forms the skeleton of the dorsal stream. Areas MT and MST are located in the deep of the superior temporal sulcus and they are considered as a pivotal hub for both object and self-motion (see, e.g., [244, 39, 246] for reviews). The motion pathway is extremely fast, with the information flowing in less that 20ms from the primary visual area to the frontal

cortices or brainstem structures underlying visuomotor transformations (see [175, 48, 209, 186] for reviews). These short time scales originate in the Magnocellular retino-geniculo-cortical input to area V1 carrying low spatial and high temporal frequencies luminance information with high contrast sensitivity (i.e., high contrast gain). This cortical input to layer 4β projects directly to the extra striate area MT, also called the cortical motion area. The fact that this feedforward stream by-passes the classical recurrent circuit between area V1 cortical layers is attractive for several reasons. First, it implements a fast, feedforward hierarchy fitting the classical two-stage motion computation model [231, 133]. Direction-selective cells in area V1 are best described as spatio-temporal filters extracting motion energy along the direction orthogonal to the luminance gradient [84, 66, 198]. Their outputs are integrated by MT cells to compute local motion direction and speed. Such spatio-temporal integration through the convergence of V1 inputs has three objectives: extracting motion signals embedded in noise with high precision, normalising them through centre-surround interactions and solving many of the input ambiguities such as the aperture and correspondance problems. As a consequence, speed and motion direction selectivities observed at single-cell and population levels in area MT are largely independent upon the contrast or the shape of the moving inputs [33, 39, 244]. The next convergence stage, area MST extracts object-motion through cells with receptive fields extending up to 10 to 20 degrees (area MSTl) or optic flow patterns (e.g., visual scene rotation or expansion) that are processed with very large receptive fields covering up to $2/3$ of the visual field (area MSTd). Second, the fast feedforward stream illustrates the fact that built-in, fast and highly specific modules of visual information are conserved through evolution to subserve automatic, behaviour-oriented visual processing (see, e.g. [209, 77, 34] for reviews). Third, this anatomical motif is a good example of a canonical circuit that implements a sequence of basic computations such as spatio-temporal filtering, gain control and normalisation at increasing spatial scales [301]. The final stage of all of these bio-inspired models consist in a population of neurons that are broadly selective for translation speed and direction [320, 253] as well as for complex optical flow patterns (see e.g., [115, 177] for recent examples). Such backbone can then be used to compute biological motion and action recognition [105, 87] similar to what was observed in human and monkey parietal cortical networks (see [106] for a recent review).

However, recent physiological studies have shown that this feedforward cornerstone of *global motion integration* must be enriched with new properties. Figure 3 depitcs some of these aspects, mirroring functional connectivity and computational perspectives. First, *motion energy estimation* through a set of spatio-temporal filters was recently re-evaluated to account for the neuronal responses to complex dynamical textures and natural images. When presented with rich, naturalistic inputs, responses of both V1 complex cells and MT pattern-motion neurons become contrast invariant [274, 73] and more selective (i.e., their tuning is sharper) [274, 104]. Their responses become also more sparse [363] and more precise [20]. These better sensitivities could be explained by a more complex integration of inputs, through a set of adaptive, excitatory- and inhibitory-weighted filters that optimally sample the spatiotemporal frequency plane [237]. Second, centre-surround interactions are much more diverse, along many different domains (e.g. retinotopic space, orientation, direction) than originally depicted by the popular Mexican-hat model. Such diversity of centre-surround interactions in both areas V1 and MT most certainly contributes to several of the computational nonlinearities mentioned above. They involve both the classical convergence of projections from one step to the next but also the dense network of lateral interactions within V1 as well as within each extra-striate areas. These lateral interactions implement long-distance normalisation, seen as centre-surround interactions at population level [285] as well as feature grouping between distant elements [107]. These intra- and inter-cortical areas interactions can support a second important aspect of motion integration: motion diffusion. In particular, anisotropic diffusion of local motion information can play a critical role

in global motion integration by propagating reliable local motion signals within the retinotopic map [340]. The exact neural implementation of these mechanisms is yet unknown but modern tools will soon allow to image, and manipulate, the dynamics of these lateral interactions. The diversity of excitatory and inhibitory inputs can explains how the aperture problem is dynamically solved by MT neurons for different types of motion inputs such as plaid patterns [301], elongated bars or barber poles [348]) and they are thought to be important to encode optic flow patterns [222] and biological motion [87]. Finally, the role of feedback in this context-dependent integration of local motion has been demonstrated by experimental [143, 234] and computational studies [21, 22] and is now addressed at the physiological level despite the considerable technical difficulties (see [72] for a review). Overall, several computational studies have shown the importance of the adaptive normalisation of spatiotemporal filters for motion perception; see [322] illustrating how a generic computation (normalisation) can be adaptively tuned to match the requirement of different behaviours.

Global motion integration is only one side of the coin. As pointed out by Braddick [38], motion integration and segmentation works hand-in-hand to selectively group the local motion signals that belong to different surfaces. For instance, some MT neurons integrate motion signals within their receptive field only if they belong to the same contour [141] or surface [329]. They can also filter out motion within the receptive field when it does not belong to the same surface [325, 329], a first step for representing motion transparency or structure-from-motion in area MT [118]. The fact that MT neurons can thus adaptively integrate local motion signals, and explain away others is strongly related to the fact that motion sensitive cells are most often embedded in *distributed multiplexed representations*. Indeed, most direction-selective cells are also sensitive to binocular disparity [176, 277, 324], eye/head motion [230] and dynamical perspective cues [159] in order to filter out motion signals from outside the plane of fixation or to disambiguate motion parallax. Thus, depth and motion processing are two intricate problems allowing the brain to compute object motion in 3D space rather than in 2D space.

Depth-motion interaction is only one example of the fact that motion pathway receives and integrates visual cues from many different processing modules [243]. This is again illustrated in Fig. 3, where form cues can be extracted in areas V2 and V4 and sent to area MT. Information about the spatial organisation of the scene using boundaries, colours, shapes might then be used to further refine the fast and coarse estimate of the optic flow that emerges from the V1-MT-MST backbone of the hierarchy. Such cue combination is critical to overcome classical pitfalls of the feedforward model. Noteworthy, along the hierarchical cascade, information is gathered over larger and larger receptive fields at the penalty that object boundaries and shapes are blurred. Thus, large receptive fields of MT and MST neurons can be useful for tracking large objects with the eyes, or avoiding approaching ones, but they certainly lower the spatial resolution of the estimated optic flow field. This feedforward, hierarchical processing contrasts with the sharp perception that we have of the moving scene. Mixing different spatial scales through recurrent connectivity between cortical areas is one solution [72, 120]. Constraining the diffusion of motion information along edges or within surface boundaries in certainly another as shown for texture-ground segmentation [307]. Such form-based representations play a significant role in disambiguation of motion information [102, 212, 210, 131]. It could also play a role in setting the balance between motion integration and segmentation dynamics, as illustrated in Fig. 3(b).

Over the last two decades, several computational vision models have been proposed to improve optic flow estimation with a bio-inspired approach. A first step is to achieve a form-independent representation of velocity from the spatio-temporal responses from V1. A dominant computational model was proposed by Heeger and Simoncelli [320], where a linear combination of afferent inputs from V1 is followed by a non linear operation known as untuned divisive normalisation. This model, and it subsequent developments [301, 237, 322] replicates a variety of observations

from physiology to psychophysics using simple, synthetic stimuli such as drifting grating and plaids. However, this class of models cannot resolve ambiguities in regions lacking of any 2D cues because of the absence of diffusion mechanisms. Moreover, their normalisation and weighted integration properties are still static. These two aspects may be the reason why they do not perform well on natural movies. Feedback signals from and to MT and higher cortical areas could play a key role in reducing these ambiguities. One good example was proposed by [21] where dynamical feedback modulation from MT to area V1 is used to solve the aperture problem locally. An extended model of V1-MT-MST interactions that uses centre-surround competition in velocity space was later presented by [281], showing good optic flow computations in the presence of transparent motion. These feedback and lateral interactions primarily play the role of context dependent diffusion operators that spread the most reliable information throughout ambiguous regions. Such diffusion mechanisms can be gated to generate anisotropic propagation, taking advantage of local form information [340, 23]. An attempt at utilising these distributed representation for integrating both optic flow estimation and segmentation was proposed in [240]. The same model explored the role of learning in establishing the best V1 representation of motion information, although this approach was largely ignored in optic flow models contrary to object categorisation for instance. In brief, more and more computational models of biological vision take advantages of these newly-elucidated dynamical properties to explain motion perception mechanisms. But it is not clear how these ideas perfuse to computer vision.

Comparison with computer vision solutions. The vast majority of computer vision solutions for optical flow estimation can be split into four major computational approaches (see [331, 93] for recent reviews). First, a constancy assumption deals with correspondence problem, assuming that brightness or color is constant across adjacent frames and assigning a cost function in case of deviation. Second, the reliability of the matching assumptions optimised using priors or a regularisation to deal with the aperture problem. Both of these solutions pose the problems as an energy function and optical flow itself is treated as an energy minimisation problem. Interestingly, a lot of recent research has been done in this area, always pushing further the limits of the state-of-the-art. This research field has put a strong emphasis on performance as a criterion to select novel approaches and sophisticated benchmarks have been developed. Since the early initiatives, current benchmarks cover a much wider variety of problems. Popular examples are the Middleburry flow evaluation [15] and, more recently the Sintel flow evaluation [51]. The later has important features which are not present in the Middlebury benchmark: long sequences, large motions, specular reflections, motion blur, defocus blur, and atmospheric effects.

Initial motion detection is a good example where biological and computer vision research have already converged. The correlation detector proposed by Hassenstein and Reichardt [121] serves as a reference for a velocity sensitive mechanisms to find correspondences of visual structure at image locations in consecutive temporal samples. Formal equivalence of correlation detection with a multi-stage motion energy filtering has been demonstrated [1]. There are now several examples of spatiotemporal filtering models that are used to extract motion energy across different scales. Initial motion detection is ambiguous since motion can locally be measured only orthogonal to an extended contrast. This is called the aperture problem and mathematically it gives an ill-posed problem to solve. For example, in gradient-based methods, one has to estimate the two velocity components from a single equation called the optical flow constraint. In spatiotemporal energy based methods, all the spatiotemporal samples lie on a straight line in frequency space and the task is to identify a plane that passes through all of them [39]. Computer vision has dealt with this problem in two ways: by imposing local constraints [194] or by posing smoothness constrains through penalty terms [140]. More recent approaches are attempted to fuse the two formulations [46]. The penalty term plays a key role as a diffusion operator can act isotropically

or anisotropically [30, 305, 12]. A variety of diffusion mechanisms has been proposed so that, e.g., optical flow discontinuities could be preserved depending on velocity field variations or image structures. All these mechanisms have demonstrated powerful results regarding the successful operation in complex scenes. Computational neurosciences models also tend to rely on diffusion mechanisms too, but they differ in their formulation. A first difference stems from the fact that local motion estimation is primarily based on the spatio-temporal energy estimation. Second, the representation is distributed, allowing multiple velocities at the same location, thus dealing with layered/transparent motion. The diffusion operator is also gated based on the local form cues also relying on the uncertainty estimate which could possibly be computed using the distributed representation [240].

Promising bio-inspired solutions. A modern trend in bio-inspired models of motion integration is to use more form-motion interactions for disambiguating information. This should be further exploited in computer vision models. Future research will have to integrate the growing knowledge about how diffusion processes, form-motion interaction and multiplexing of different cues are implemented and impact global motion computation [348, 280, 213]. Despite the similarities in the biological and artificial approaches to solve optical flow computation, it is important to note that there is only little interaction happening between computer vision engineers and biological vision modellers. One reason might be that biological models have not been rigorously tested on regular computer vision datasets and are therefore considered as specifically confined to laboratory conditions only. It would thus be very interesting to evaluate models such as [320, 22, 42, 339] to identify complementary strengths and weaknesses in order to find converging lines of research investigations. Figure 8 illustrates work initiated in this direction where three bio-inspired models that have been tested on the Middlebury optical flow dataset [15]. Each of these models describe a potential strategy applied by the biological visual systems to solve motion estimation problem. The first model [326], demonstrates the applicability of a feedforward model that has been suggested for motion integration by MT neurons [301] for estimation of optical flow by extending it into a scale-space framework and applying a linear decoding scheme for conversion of MT population activity into velocity vectors. The second model [215] investigates the role of contextual adaptations depending on form based cues in feedforward pooling by MT neurons. The third model [37] studies the role of modulatory feedback mechanisms in solving the aperture problem.

Some elements of the mechanisms discussed above (e.g. the early motion detection stage, [125]) have already been incorporated in recent computer vision models, For instance, the solution proposed by [370] uses a regularisation scheme that considers different temporal scales, namely a regular motion mechanism (using short exposure frames) as well as a slowly integrating representation (using long exposure frames), the latter resembling the form pathway in the primate visual system [308]. The goal there was to reduce inherent uncertainty in the input [197]. Further constraining the computer vision models by simultaneously including some of the above-described mechanisms (e.g. tuned normalisation through lateral interactions, gated pooling to avoid estimation errors, feedback-based long range diffusion) may lead to significant improvements in optic flow processing methods and engineering solutions.

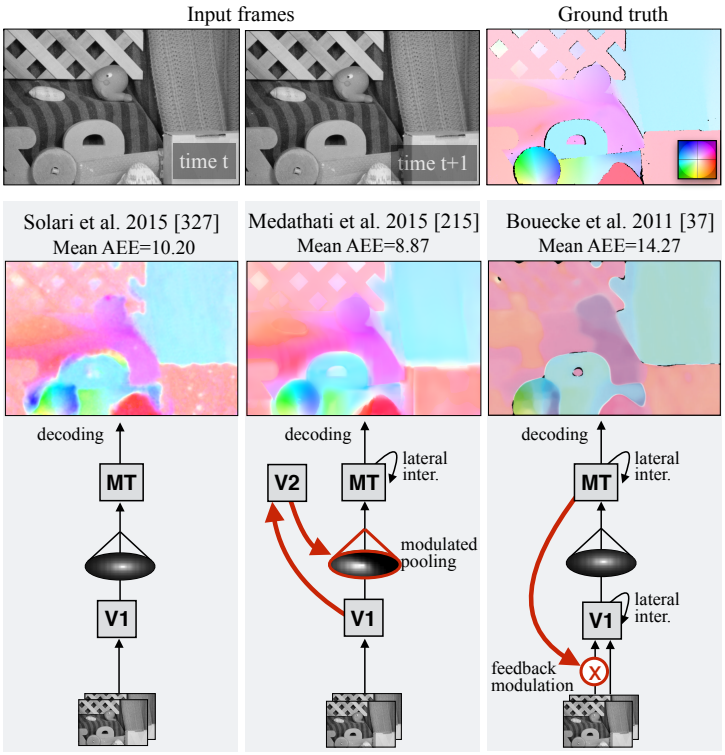


Figure 8: Comparison between three biological vision models tested on the Rubberwhale sequence from Middlebury dataset [15]. First column illustrates [326], where the authors have revisited the seminal work by Heeger and Simoncelli [320] using spatio-temporal filters to estimate optical flow from V1-MT feedforward interactions. Second column illustrates [215], an extension of the Heeger and Simoncelli model with adaptive processing algorithm based on context-dependent, area V2 modulation onto the pooling of V1 inputs onto MT cells. Third column illustrates [37], which incorporates modulatory feedbacks from MT to V1. Optical flow is represented using the colour-code from Middlebury dataset.

5 Discussion

In Sec. 4 we have revisited three classical computer vision tasks and discussed strategies that seemed to be used by biological vision systems in order to solve them. Tables 1 and 2 provide a concise summary of existing models for each task, together with key references about corresponding biological findings. From this meta-analysis, we have identified several research flows from biological vision that should be leveraged in order to advance computer vision algorithms. In this section, we will briefly discuss some of the major theoretical aspects and challenges described throughout the review.

5.1 Structural principles that relate to function

Studies in biological vision reveal structural regularities in various regions of the visual cortex. For decades, the hierarchical architecture of cortical processing has dominated, where response selectivities become more and more elaborated across levels along the hierarchy. The potential for using such deep feedforward architectures for computer vision has recently been discussed by [169]. However, it appears nowadays that such principles of bottom-up cascading should be combined with lateral interactions within the different cortical functional maps and the massive feedback from higher stages. We have indicated several computations (e.g., normalisation, gain control, segregation...) that could be implemented within and across functional maps by these connectivity motives. We have shown the impact of these interactions on each of the three example tasks (sensing, segmentation, optic flow) discussed throughout this article. We have also mentioned how these bio-inspired computational blocks (e.g., normalisation) can be re-used in a computer vision framework to improve image processing algorithms (e.g., statistical whitening and source separation [196], pattern recognition [148]). One fundamental aspect of lateral and feedback interactions is that they implement context-dependent tuning of neuronal processing, over short distance (e.g. the classical centre-surround interactions) but also over much larger distances (e.g. anisotropic diffusion, feature-based attention). We have discussed the emerging ideas that these intricate, highly recurrent architectures are key ingredients to obtain an highly-flexible visual system that can be dynamically tuned to the statistics of each visual scene and to the demands of the on-going behavioural task on a moment-by-moment basis. It becomes indispensable to better understand and model how these structural principles, for which we are gaining more and more information every day, relate to functional principles. What is important in sensing, segmenting and computing optical flow is not much what could be the specific receptive fields involved in each of these problems but, rather to identify the common structural and computational architectures that they share (see Box 1). For instance, bottom-up signal representations and top-down predictions would achieve a resonant state in which the context re-enters the earlier stages of representation in order to emphasise their relevance in a larger context [112, 80]. These interactions are rooted in the generic mechanisms of response normalisation based on non-linear divisive processes. A corresponding canonical circuit, using spiking neurons representations, can then be proposed, as in [43] for instance. Variants of such computational elements have been used in models tackling each of these three example task; sensing, segmenting and optical flow (e.g., [21, 22, 375, 340]) using either functional models or neural fields formalism (see Box 1). More important, these different models can be tested on a set of real-world images and sequences taken from computer vision. This is just one exemple of the many different instances of operative solutions and algorithms that can be inspired from biology and computational vision. It is important to consider that the computational properties of a given architecture (e.g. recurrent connectivity) have been investigated in different theoretical perspectives (e.g, Kalman filtering) and different mathematical frameworks (e.g., [279, 78, 252]).

	Reference	Model	Application	Code
Sensing	Vanrullen et al., 2002 [359]	Spatial model based on difference-of-Gaussian kernels at different scales	Object recognition using the idea of latency coding	○
	Benoit et. al., 2010 [26]	Spatio-temporal model of retinal parvocellular and magnocellular pathways (also includes a V1 model)	Low level image processing	●
	Wohrer et al., 2009 [375]	Spiking retina model with contrast gain control (*Virtual Retina*)	Comparisons to single cell recordings and large scale simulations	●
	Lorach et al., 2012, [192]	Retina-inspired sensor combining an asynchronous event-based light sensor (DVS) with a model pulling non-linear subunits to reproduce the parallel filtering and temporal coding of the majority of ganglion cell types	Target artificial visual systems and visual prosthetic devices	○
	Martinez et al., 2013, [205]	Compiler-based framework with an ad hoc language allowing to produce accelerated versions of the models compatible with COTS microprocessors, FPGAs or GPUs (*Retina Studio*)	Target visual prosthetic devices	○
Segmentation	Parent et al., 1989 [249]	Model of curve detection and boundary grouping using tangent orientation and local curvature information	Tested on artificial noisy images for curve evaluation and natural images from different domains	○
	Ren et al., 2006 [284]	Figure-ground assignment to contours in natural images based on mid-level visual shapes (so-called shapemes) and global consistency enforcement for contour junctions	Bottom-up figure-ground label assignment in still images of large data bases with human ground truth labellings	○
	Bornstein et al., 2008 [32]	Model for image segmentation combining bottom-up processing (to create hierarchies of segmented uniform regions) with top-down processing (to employ shape knowledge from prior learning of image fragments)	Tested on data sets with four classes of objects to demonstrate improved segmentation and recognition performance	○
	Rodriguez et al., 2012 [289]	Computational model of mid-level 2D shape representation utilizing hierarchical processing with end-stopping and curvature selective cells	Tested on artificial shape configurations to replicate experimental findings from neurophysiology	○
	Azzopardi et al., 2012 [14]	Computational model of center-surround and orientation selective filtering with non-linear context-dependent suppressive modulation and cross-orientation inhibition	Tested on two public data sets of natural images with contour ground truth labellings	○
	Tschechne, 2014 [341]	Recurrent network architecture for distributed multi-scale shape feature representation, boundary grouping, and border-ownership direction assignment	Tested on a selection of stimuli from public data sets	○
Optical flow	Heeger, 1988 [125]	Feed forward model based on spatio-temporal motion energy filters	Used to simulate psychophysical data and Yosemite sequence	○
	Nolan et al., 1994 [240]	Model based on spatio-temporal motion energy filters with a selection mechanism to deal with occlusions and transparency	Optical flow estimation, tested on synthetic images only	○
	Grossberg et al., 2001 [117]	Dynamical model representative of interactions between V1, V2, MT and MST areas	Grouping and optical flow estimation, tested on synthetic images only	○
	Bayerl et al., 2007 [22]	Recurrent model of V1-MT with modulatory feedbacks and a sparse coding framework for neural motion activity patterns	Optical flow estimation, tested using several real world classical videos	○
	Tlapale et al., 2010 [340]	Dynamical model representative of V1-MT interactions and luminosity based motion information diffusion	Optical flow estimation, tested on synthetic images only	○
	Perrone et al., 2012 [253]	Model explaining the speed tuning properties of MST neurons by afferent pooling from MT	Optical flow estimation, tested on synthetic and two natural sequences	○
	Tschechne et al., 2014 [342]	Model of cortical mechanisms of motion detection using an asynchronous event-based light sensor (DVS)	Motion estimation with limited testing for action recognition	○
	Solari et al., 2015 [326] RR n° 8698	Multi-scale implementation of a feedforward model based on spatio-temporal motion energy filters inspired by [125]	Dense optical flow estimation, evaluated on Middlebury benchmark	●

Table 1: Highlight on models for each of the three tasks considered in Sec. 4.

	Biological mechanism	Experimental paper	Models
Sensing	Visual adaptation	[314, 336, 156]	[375, 129]
	Feature detection	[156]	[129]
	Sparse coding	[260]	[192]
	Precision	[260]	[192]
	Surveys	[206, 207]	–
Segmentation	Contrast enhancement and shape representation	[101]	[14, 289]
	Feature integration and segmentation	[47, 255, 91, 36, 155, 318, 376, 182, 107]	[114, 116, 204, 235, 24, 53, 32, 11]
	Border ownership and figure-ground segregation	[172, 174, 143, 385, 256, 150, 307, 381]	[113, 284, 69, 94, 137, 341]
	Continuation and visual routines	[151, 157, 267, 164]	[122, 283]
	Surveys	-	[130, 26, 68]
Optical flow	Motion energy estimation	[84, 66, 198, 302]	[1, 125, 320]
	Local velocity estimation	[335, 301, 275, 39, 237]	[237, 326]
	Global motion integration	[141]	[240, 117, 22, 340, 253]
	Distributed multiplexed representations	[211, 18, 230, 142, 324]	[49, 176, 277, 89, 243]
	Surveys	[246, 231]	[37]

Table 2: Summary of the strategies highlighted in the text to solve the different task, showing where to find more details about the biological mechanisms and which models are using these strategies.

Some of the biologically-plausible models assembled in Tables 1 offer a repertoire of realistic computational solutions that can be a source of inspiration for novel computer vision algorithms.

5.2 Data encoding and representation

Biological systems are known to use several strategies such as event-based sensory processing, distributed multiplexed representation of sensory inputs and active sensory adaptation to the input statistics in order to operate in a robust and energy efficient manner. Traditionally, video inputs are captured by cameras that generate sequences of frames at a fixed rate. The consequence is that the stream of spatio-temporal scene structure is regularly sampled at fixed time steps regardless of the spatio-temporal structure. In other words, the plenoptic function [2] is sliced in sheets of image-like representations. The result of such a strategy is a highly redundant representation of any constant features in the scene along the temporal axis. In contrast, the brain encodes and transmits information through discrete sparse events and this spiking encoding appears at the very beginning of visual information processing, i.e., at the retina level. As discussed in Sec. 4.1, ganglion cells transmit a sparse asynchronous encoding of the time varying visual information to LGN and then cortical areas. This sparse event-based encoding inspired development of new type of camera sensors. Some events are registered whenever changes occur in the spatio-temporal luminance functions which are represented in a stream of events, with a location and time stamp [184, 188, 269]. Apart from the decrease in redundancy, the processing speed is no longer restricted to the frame-rate of the sensor. Rather, events can be delivered at a rate that is only limited by the refractory period of the sensor elements. Using these sensors

brings massive improvements in terms of efficiency of scene encoding and computer vision approaches could benefit from such an alternative representation as demonstrated already on some isolated tasks.

In terms of representation, examining the richness of receptive fields of cells from retina of the visual cortex (such as in V1, MT and MST) shows that the visual system is almost always using a distributed representation for the sensory inputs. Distributed representation helps the system in a multiplicity of ways: It allows for an inherent representation for the uncertainty, it allows for task specific modulation and it could also be useful for representing the multiplicity of properties such as transparent/layered motion [272, 321]. Another important property of biological vision that visual features are optimally encoded at the earliest stages for carrying out computations related to multiplicity of tasks in higher areas. Lastly, we have briefly mentioned that there are several codes to be used by visual networks in order to represent the complexity of natural visual scenes. Thus, it shall be very helpful to take into account this richness of representations to design systems that could deal with an ensemble of tasks simultaneously instead of subserving a single task at a time.

Recently, the application of DCNNs to solve computer vision tasks has boosted machine performance in processing complex scenes, achieving human level performance in certain scenarios. Their hierarchical structure and the utilisation of simple canonical operations (filtering, pooling, normalisation, etc.) motivated investigators to test their effectiveness in predicting cortical cell responses [262, 119]. In order to generate artificial networks with functional properties which come close to primate cortical mechanisms, a goal-diven modelling approach has been proposed which achieved promising results [380]. Here, the top-layer representations should be constrained in the learning by the particular task of the whole network. The implicit assumption is that such a definition of the computational goal lies in the overlapping region of artificial and human vision systems, since otherwise the computational goals might deviate between systems as discussed above [346] (his Fig.1). The authors argue that the detailed internal structures might deviate from those identified in cortex, but additional auxiliary optimisation mechanisms might be employed to vary structures under the constraint to match the considered cortical reference system [28]. The rating of any network necessitates the definition of a proper similarity measure, such as using dissimilarity measures computed from response patterns of brain regions and model representations to compare the quality of the input stimulus representations [166].

5.3 Psychophysics and human perceptual performance data

Psychophysical laws and principles which can explain large amounts of empirical observations should be further explored and exploited for designing robust vision algorithms. However, most of our knowledge about human perception has been gained using either highly artificial inputs for which the information is well-defined or natural images for which the information content is much less known. By contrast, human perception continuously adjusts information processing to the content of the images, at multiple scales and depending upon different brain states such as attention or cognition. For instance, human vision dynamically tuned decision-boundaries related to changes observed in the environment. It has been demonstrated that this adaptation can be achieved dynamically by non-linear network properties that incorporate activation transfer functions of sigmoidal shape [112]. In [61], such a principle has been adopted to define a robust image descriptor that adjusts its sensitivity to the overall signal energy, similar to human sensitivity shifts. One of the fondamental advantages of these formalism is that they can render the biological performance at many different levels, from neuronal dynamics to human performance. In other words, they can be used to adjust the algorithm parameters to different levels of constraints shared by both biological and computer vision [346]

Most of the problems in computer vision are ill-posed and observable data are insufficient in terms of variables to be estimated. In order to overcome this limitation, biological systems exploit statistical regularities. The data from human performance studies either on highly controlled stimuli with careful variations in specific attributes or large amounts of unstructured data can be used to identify the statistical regularities, particularly significant for identifying operational parameter regimes for computer vision algorithms. This strategy is already being explored in computer vision and is becoming more popular with the introduction of larges scale internet based labelling tools such as [300, 367, 349]. Classic examples for this approach in the case of scene segmentation are exploration of human marked ground truth data for static [204] and dynamic scenes [100]. Thus, we advocate that further investigation on the front-end interfaces to learning functions, decision-making or separation boundaries for classifiers might improve the performance levels of existing algorithms as well as their next generations. Emerging work such as [304] illustrates the potential in this direction. [304] use the human performance errors and difficulties for the task of face detection to bias the cost function of the SVM to get closer to the strategies that we might be adapting or trade-offs that our visual systems are banking on. We have provided other examples throughout the article but it is evident that further linking learning approaches with low- and mid-levels of visual information is a source of major advances in both understanding of biological vision and designing better computer vision algorithms.

5.4 Computational models of cortical processing

Over the last decade, many computational models have been proposed to give a formal description of phenomenological observations (e.g. perceptual decisions, population dynamics) as well as a functional description of identified circuits. Throughout this article, we have proposed that bio-inspired computer vision shall consider the existence of a few generic computational modules together with their circuit implementation. Implementing and testing these canonical operations is important to understand how efficient visual processing as well as highly flexible, task-dependent solutions can be achieved using biological circuit mechanisms and and to implement them within artificial systems. Moreover, the genericness of visual processing systems can be viewed as an emergent property from an appropriate assembly of these canonical computational blocks within a dense, highly recurrent neural networks. Computational neurosciences also investigate the nature of the representations used by these computational blocks (e.g., probabilistic population codes, population dynamics, neural maps) and we have proposed how such new theoretical ideas about neural coding can be fruitful to move forward beyond the classical isolated processing units that are typically approximated as linear-non linear filters. For each of the three example tasks, we have indicated several computational operative solutions that can be inspiring for computer vision. Table 1 highlights a selection of papers where even a large panels of operative solutions are described. It is beyond the scope of this paper to provide a detailed mathematical framework for each problem described or a comprehensive list of operative solutions. Still, in order to illustrate our approach, we provide in Box 1 three examples of popular operative solutions that can translate from computational to computer vision. These three examples are representative of the different mathematical frameworks described above: a functional model such as divisive normalisation that can be used for regulating population coding and decoding; a population dynamics model such as neural fields that can be used for coarse level description of lateral and feedback interactions and, lastly a neuromorphic representation data and of event-based computations such as spiking neuronal models.

The field of computational neurosciences has made enormous progress over the last decades and will be boosted by the flow of new data gathered at multiple scales, from behaviour to synapses. Testing popular computational vision models against classical benchmarks in computer

vision is a first step needed to bring together these two fields of research, as illustrated above for motion processing. Translating new theoretical ideas about brain computations to artificial systems is a promising source of inspiration for computer vision as well. Both computational and computer vision share the same challenge: each one is the missing link between hardware and behaviour, in search for generic, versatile and flexible architectures. The goal of this review was to propose some aspects of biological visual processing for which we have enough information and models to build these new architectures.

Box 1 | Three examples of operative solutions

Normalization is a generic operation present at each level of the visual processing flow, playing critical role in functions such as controlling contrast gain or tuning response selectivity [57]. In the context of neuronal processing, the normalization of the response R_i of a single neuron can be written by

$$R_i = \frac{{I_i}^n}{k_{tuned}{I_i}^n + \sum_j W_{ij}(I_j)^n + \sigma},$$

where $I_{\{.\}}$ indicates the net excitatory input to the neuron, $(\sum_j)$ indicates the summation over normalization pool, σ is a stabilization constant, W_{ij} are weights, n and k_{tuned} are the key parameters regulating the behavior. When $k_{tuned} = 0$ and $n = 1$ this equation represents a standard normalization. When the constant k_{tuned} is non-zero, normalization is referred to as tuned normalization. This notion has been used in computational models for, e.g., tone mapping [219] or optical flow [21, 326].

The dynamics of biological vision results from the interaction between different cortical streams operating at different speeds but also relies on a dense network of intra-cortical and inter-cortical connections. Dynamics is generally modelled by neural fields equations which are spatially structured neural networks which represent the spatial organization of cerebral cortex [40]. For example, to model the dynamics of two populations $p_1(t,r)$ and $p_2(t,r)$ (where $p.$ is the firing activity of each neural mass and r can be thought of as defining the population), a typical neural field model is

$$\frac{\partial p_1}{\partial t} = \underbrace{-\lambda_1 p_1}_{\text{decay}} + S\Big(\int_{r'} \underbrace{W_{1\to1}(t,r,r')}_{\text{lateral}} p_1(t,r') + \int_{r'} \underbrace{W_{2\to1}(t,r,r')}_{\text{feedback}} p_2(t,r') + \underbrace{K(t,r)}_{\text{external input}} \Big),$$

$$\frac{\partial p_2}{\partial t} = \underbrace{-\lambda_2 p_2}_{\text{decay}} + S\Big(\int_{r'} \underbrace{W_{1\to2}(t,r,r')}_{\text{feedforward}} p_1(t,r') + \int_{r'} \underbrace{W_{2\to2}(t,r,r')}_{\text{lateral}} p_2(t,r') \Big),$$

where the weights $W_{i\to j}$ represent the key information defining the connectivities and $S(\cdot)$ is a sigmoïdal function. Some example of neural fields model in the context of motion estimation are [340, 339, 278].

Event driven processing is the basis of neural computation. A variety of equations have been proposed to model the spiking activity of single cells with different degrees of fidelity to biology [103]. A simple classical case is the leaky-integrate and fire neuron (seen as a simple RC circuit) where the membrane potential u_i is given by

$$\tau \frac{du_i}{dt} = -u_i(t) + RI_t(t),$$

with a spike emission process: the neuron i will emit a spike when $u_i(t)$ reaches a certain threshold. τ is time constant of the leaky integrator and R is the resistance of the neuron. When the neuron belongs to a network, the input current is given by $I_i(t) = \sum_j W_{j\to i} \sum_f \alpha(t - t_j^{(f)})$ where $t_j^{(f)}$ represents the time of the f-th spike of the j-th pre-synaptic neuron, $\alpha(t)$ represents the post synaptic current generated by the spike and $W_{j\to i}$ is the strength of the synaptic efficacy from neuron j to neuron i. This constitutes the building block of a spiking neural network. In term of neuromorphic architectures, this principle has inspired sensors such as event-based cameras (see Sec. 4.1). From a computation point of view, it has been used for biological vision [375, 192] but also for solving vision tasks [88, 208].

6 Conclusion

Computational models of biological vision aim at identifying and understanding the strategies used by visual systems to solve problems which are often the same as the one encountered in computer vision. As a consequence, these models would not only shed light into functioning of biological vision but also provide innovative solutions to engineering problems tackled by computer vision. In the past, these models were often limited and able to capture observations at a scale not directly relevant to solve tasks of interest for computer vision. More recently, enormous advances have been made by the two communities. Biological vision is quickly moving towards systems level understanding while computer vision has developed a great deal of task centric algorithms and datasets enabling rapid evaluation. However, computer vision engineers often ignore ideas that are not thoroughly evaluated on established datasets and modellers often limit themselves to evaluating highly selected set of stimuli. We have argued that the definition of common benchmarks will be critical to compare biological and artificial solutions as well as integrating recent advances in computational vision into new algorithms for computer vision tasks. Moreover, the identification of elementary computing blocks in biological systems and their interactions within highly recurrent networks could help resolving the conflict between task-based and generic approach of visual processing. These bio-inspired solutions could help scaling up artificial systems and improve their generalisation, their fault-tolerance and adaptability. Lastly, we have illustrated how the richness of population codes, together with some of their key properties such as sparseness, reliability and efficiency could be a fruitful source of inspiration for better representations of visual information. Overall, we argue in this review that despite their recent success, machine vision shall turn the head again towards biological vision as a source of inspiration.

Acknowledgements

NVK.M. acknowledges support from the EC IP project FP7-ICT-2011-8 no. 318723 (MathemACS) H.N. acknowledges support from DFG in the SFB/TRR 'A Companion Technology for Cognitive Technical Systems. P.K. acknowledges support from the EC IP project FP7-ICT-2011-9 no. 600847 (RENVISION). GSM is supported by the European Union (Brainscales, FP7-FET-2010-269921), the french ANR (SPEED, ANR-13-SHS2-0006) and CNRS.

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2:284–299, 1985. 37, 42

[2] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991. 42

[3] M. Ahissar and S. Hochstein. The reverse hierarchy of visual perceptual learning. *Trends in Cognitive Sciences*, 8(10):457–464, 2004. 11

[4] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 510—-517, 2012. 25

[5] B. Alexander, H. Juergen, and F. R. Dierk. Fly motion vision. *Annual Review of Neuroscience*, 33(1):49–70, 2010. PMID: 20225934. 34

[6] I. Andolina, H. Jones, W. Wang, and A. Sillito. Corticothalamic feedback enhances stimulus response precision in the visual system. *Proceedings of the National Academy of Sciences*, 104(1685–1690), 2007. 9

[7] A. Andreopoulos and J. K. Tsotsos. 50 years of object recognition: Directions forward. *Computer Vision and Image Understanding*, 117:827–891, 2013. 15, 21

[8] A. Angelucci and P. C. Bressloff. Contribution of feedforward, lateral and feedback connections ot the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. *Progress in Brain Research*, 154:93–120, 2006. 12

[9] A. Angelucci and J. Bullier. Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons? *Journal of Physiology - Paris*, 97(2–3):141–154, 2003. 17

[10] A. Anzai, X. Peng, and D. Van Essen. Neurons in monkey visual area V2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–1321, 2007. 29

[11] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, May 2011. 28, 31, 42

[12] G. Aubert and P. Kornprobst. *Mathematical problems in image processing: partial differential equations and the calculus of variations (Second edition)*, volume 147 of *Applied Mathematical Sciences*. Springer-Verlag, 2006. 38

[13] V. Auvray, P. Bouthemy, and J. Liénard. Joint Motion Estimation and Layer Segmentation in Transparent Image Sequences—Application to Noise Reduction in X-Ray Image Sequences. *EURASIP Journal on Advances in Signal Processing*, 2009, 2009. 34

[14] G. Azzopardi and N. Petkov. A CORF computational model of a simple cell that relies on LGN input outperforms the gabor function model. *Biological Cybernetics*, 106(3):177–189, 2012. 41, 42

[15] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011. 19, 37, 38, 39

[16] D. Ballard, M. Hayhoe, G. Salgian, and H. Shinoda. Spatio-temporal organization of behavior. *Spatial Vision*, 13(2-3):321–333, 2000. 26

[17] D. H. Ballard and M. M. HayHoe. Modelling the role of task in the control of gaze. *Visual Cognition*, 17:1185–1204, 2009. 32

[18] A. Basole, L. White, and D. Fitzpatrick. Mapping multiple features in the population response of visual cortex. *Nature*, 423:986–990, 2003. 42

[19] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695 – 711, 2012. 16

[20] P. Baudot, M. Levy, O. Marre, M. Pananceau, and Y. Fregnac. Animation of natural scene by virtual eye-movements evoke high precision and low noise in V1 neurons. *Frontiers in Neural Circuits*, 7:206, 2013. 35

[21] P. Bayerl and H. Neumann. Disambiguating visual motion through contextual feedback modulation. *Neural Computation*, 16(10):2041–2066, 2004. 12, 36, 37, 40, 46

[22] P. Bayerl and H. Neumann. Disambiguating visual motion by form–motion interaction – a computational model. *International Journal of Computer Vision*, 72(1):27–45, 2007. 36, 38, 40, 41, 42

[23] C. Beck and H. Neumann. Interactions of motion and form in visual cortex – a neural model. *Journal of Physiology - Paris*, 104:61–70, 2010. 18, 37

[24] O. Ben-Shahar and S. Zucker. Geometrical computations explain projection patterns of long-range horizontal connections in visual cortex. *Neural Computation*, 16(3):445—476, 2004. 31, 42

[25] A. Benoit, D. Alleysson, J. Hérault, and P. Le Callet. Spatio-temporal tone mapping operator based on a retina model. In *Computational Color Imaging Workshop*, 2009. 25

[26] A. Benoit, A. Caplier, B. Durette, and J. Herault. Using human visual system modeling for bio-inspired low level image processing. *Computer Vision and Image Understanding*, 114(7):758 – 773, 2010. 24, 26, 41, 42

[27] R. Benosman, S.-H. Ieng, C. Clercq, C. Bartolozzi, and M. Srinivasan. Asynchronous frameless event-based optical flow. *Neural Networks*, 27:32–37, 2011. 26

[28] J. Bergstra, D. Yamins, and D. Cox. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, pages 115–123, Atlanta, Georgia, USA, June 2013. 19, 43

[29] M. Bertalmío. *Image Processing for Cinema*. CRC Press, 2014. 25

[30] M. Black, G. Sapiro, D. Marimont, and D. Heeger. Robust anisotropic diffusion. *IEEE Transactions on Image Processing*, 7(3):421–432, 1998. Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing and Analysis. 38

[31] D. Bock, W.-C. A. Lee, A. Kerlin, M. Andermann, G. Hood, A. Wetzel, S. Yurgenson, E. Soucy, H. Kim, and R. Reid. Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, 471(177-182), 2011. 5

[32] E. Borenstein and S. Ullman. Combined top-down/bottom-up segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2109–2125, 2008. 26, 31, 41, 42

[33] R. Born and D. Bradley. Structure and function of visual area MT. *Annual Review of Neuroscience*, 28:157–189, 2005. 12, 35

[34] A. Borst. Fly visual course control: behaviour, algorithms and circuits. *Nature Reviews Neuroscience*, 15:590–599, 2014. 34, 35

[35] A. Borst and T. Euler. Seeing things in motion: Models, circuits, and mechanisms. *Neuron*, 71(6):974 – 994, 2011. 34

[36] W. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *The Journal of Neuroscience*, 17(6):2112–2127, 1997. 29, 42

[37] J. Bouecke, E. Tlapale, P. Kornprobst, and H. Neumann. Neural mechanisms of motion detection, integration, and segregation: From biology to artificial image processing systems. *EURASIP Journal on Advances in Signal Processing*, 2011. Special issue on Biologically inspired signal processing: Analysis, algorithms, and applications. 18, 38, 39, 42

[38] O. Braddick. Segmentation versus integration in visual motion processing. *Trends in neurosciences*, 16(7):263–268, 1993. 17, 33, 36

[39] D. Bradley and M. Goyal. Velocity computation in the primate visual system. *Nature Reviews Neuroscience*, 9(9):686–695, 2008. 7, 34, 35, 37, 42

[40] P. Bressloff. Spatiotemporal dynamics of continuum neural fields. *Journal of Physics A: Mathematical and Theoretical*, 45, 2012. 46

[41] F. Briggs and W. M. Usrey. Emerging views of corticothalamic function. *Current Opinion in Neurobiology*, 18(4):403–407, Aug. 2008. 9

[42] R. S. A. Brinkworth and D. C. Carroll. Robust models for optic flow coding in natural scenes inspired by insect biology. *PLoS Computational Biology*, 5(11), 2009. 38

[43] T. Brosch and H. Neumann. Interaction of feedforward and feedback streams in visual cortex in a firing-rate model of columnar computations. *Neural Networks*, 54(0):11 – 16, 2014. 40

[44] T. Brosch, H. Neumann, and P. Roelfsema. Reinforcement learning of linking and tracing contours in recurrent neural networks. *PLoS Comput Biol*, 11(10), 2015. 12

[45] T. Brosch, S. Tschechne, and H. Neumann. On event-based optical flow detection. *Frontiers in Neuroscience*, 9(137), Apr. 2015. 26

[46] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61:211–231, 2005. 38

[47] E. Brunswik and J. Kamiya. Ecological cue-validity of 'proximity' and of other gestalt factors. *The American Journal of Psychology*, 66(1):20–32, 1953. 29, 42

[48] J. Bullier. Integrated model of visual processing. *Brain Res. Reviews*, 36:96–107, 2001. 11, 13, 35

[49] G. T. Buracas and T. D. Albright. Contribution of area MT to perception of three-dimensional shape: a computational study. *Vision Res*, 36(6):869–87, 1996. 42

[50] T. Buschman and S. Kastner. From behavior to neural dynamics: an integrated theory of attention. *Neuron*, 88(7):127–144, 2015. 12, 13

[51] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the 12th European Conference on Computer Vision*, pages 611–625. Springer-Verlag, 2012. 19, 37

[52] Z. Bylinskii, E. DeGennaro, H. Rajalingham, R.and Ruda, J. Zhang, and J. Tsotsos. Towards the quantitative evaluation of visual attention models. *Vision Research*, 116:258–268, 2015. 12

[53] C. Cadieu, M. Kouh, A. Pasupathy, C. Connor, M. Riesenhuber, and T. Poggio. A model of V4 shape selectivity and invariance. *Journal of Neurophysiology*, 98(1733-1750), 2007. 8, 42

[54] J. F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):769–798, Nov. 1986. 30

[55] M. Carandini. From circuits to behavior: a bridge too far? *Nature Publishing Group*, 15(4):507–509, Apr. 2012. 15

[56] M. Carandini, J. B. Demb, V. Mante, D. J. Tollhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597, Nov. 2005. 24

[57] M. Carandini and D. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2011. 13, 16, 20, 46

[58] J. R. Cavanaugh, W. Bair, and J. A. Movshon. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *Journal of Neurophysiology*, 88(5):2530–2546, 2002. 13

[59] L. M. Chalupa and J. Werner, editors. *The visual neurosciences*. MIT Press, 2004. Two volumes. 5

[60] R. Chaudhuri, K. Knoblauch, M. A. Gariel, K. H, and X. J. Wang. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron*, 88:419–431, 2015. 5, 10

[61] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao. WLD: a robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720, September 2010. 43

[62] E. J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Computatiton in Neural Systems*, 12:199–213, 2001. 24

[63] G. Citti and A. Sarti, editors. *Neuromathematics of Vision*. Springer, 2014. 32

[64] M. Cohen and J. Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, 12:1594–1600, 2009. 12

[65] G. Coleman and H. C. Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979. 30

[66] B. Conway and M. Livingstone. Space-time maps and two-bar interactions of different classes of direction-selective cells in macaque V1. *Journal of Neurophysiology*, 89:2726–2742, 2003. 35, 42

[67] D. D. Cox. Do we understand high-level vision? *Current Opinion in Neurobiology*, 25(187-193), 2014. 6

[68] D. D. Cox and T. Dean. Neural networks and neuroscience-inspired computer vision. *Current Biology*, 24(18):921–929, 2014. 6, 8, 20, 42

[69] E. Craft, H. Schutze, E. Niebur, and R. von der Heydt. A neural model of figure-ground organization. *Journal of Neurophysiology*, 97:4310–4326, 2007. 32, 42

[70] G. Cristóbal, L. Perrinet, and M. S. Keil, editors. *Biologically Inspired Computer Vision: Fundamentals and Applications*. Wiley-VCH, 2015. 6, 15

[71] A. Cruz-Martin, R. El-Danaf, F. Osakada, B. Sriram, O. Dhande, P. Nguyen, E. Callaway, A. Ghosh, and A. Huberman. A dedicated circuit links direction-selective retinal ganglion cells to the primary visual cortex. *Nature*, 507:358–361, 2014. 16

[72] J. Cudeiro and A. M. Sillito. Looking back: corticothalamic feedback and early visual processing. *Trends in Neurosciences*, 29(6):298–306, June 2006. 9, 12, 36

[73] Y. Cui, L. Liu, F. Khawaja, C. Pack, and D. Butts. Diverse suppressive influences in area MT and selectivity to complex motion features. *Journal of Neuroscience*, 33(42):16715–16728, 2013. 35

[74] J. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1169–1179, 1988. 15

[75] J. B. Demb. Functional circuitry of visual adaptation in the retina. *The Journal of Physiology*, 586(18):4377–4384, 2008. 22

[76] E. DeYoe and D. C. Van Essen. Concurrent processing streams in monkey visual cortex. *Trends in Neurosciences*, 11(219-226), 1988. 7, 11

[77] O. Dhande and A. Huberman. Retinal ganglion cell maps in the brain: implications for visual processing. *Current Opinion in Neurobiology*, 24:133–142, 2014. 35

[78] K. Dimova and M. Denham. A neurally plausible model of the dynamics of motion integration in smooth eye pursuit based on recursive bayesian estimation. *Biological Cybernetics*, 100(3):185–201, 2009. 16, 40

[79] R. Eckhorn, H. Reitboeck, M. Arndt, and P. Dicke. Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex. *Neural Computation*, 2(3):293–307, 1990. 13

[80] G. M. Edelman. Neural darwinism: Selection and reentrant signaling in higher brain function. *Neuron*, 10(2):115 – 125, 1993. 29, 40

[81] Editorial. Focus on neurotechniques. *Nature Neuroscience*, 16(7):771–771, June 2013. 5

[82] H. Eichner, T. Klug, and A. Borst. Neural simulations on multi-core architectures. *Frontiers in Neuroinformatics*, 3(21), 2009. 5

[83] G. Eilertsen, R. Wanat, R. Mantiuk, and J. Unger. Evaluation of tone mapping operators for HDR-Video. *Computer Graphics Forum*, 32(7):275–284, Oct. 2013. 25

[84] R. Emerson, J. Bergen, and E. Adelson. Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Research*, 32:203–218, 1992. 35, 42

[85] A. K. Engel and W. Singer. Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences*, 5(1):16 – 25, 2001. 13

[86] C. Enroth-Cugell and J. Robson. Functional characteristics and diversity of cat retinal ganglion cells. basic characteristics and quantitative description. *Investigative Ophthalmology and Visual Science*, 25(250-257), 1984. 9

[87] M.-J. Escobar and P. Kornprobst. Action recognition via bio-inspired features: The richness of center-surround interaction. *Computer Vision and Image Understanding*, 116(5):593—605, 2012. 35, 36

[88] M.-J. Escobar, G. S. Masson, T. Viéville, and P. Kornprobst. Action recognition using a bio-inspired feedforward spiking network. *International Journal of Computer Vision*, 82(3):284, 2009. 46

[89] J. Fernandez, B. Watson, and N. Qian. Computing relief structure from motion with a distributed velocity and disparity representation. *Vision Research*, 42(7):863–898, 2002. 42

[90] S. Ferradans, M. Bertalmio, E. Provenzi, and V. Caselles. An Analysis of Visual Adaptation and Contrast Perception for Tone Mapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2002–2012, Oct. 2011. 25

[91] D. Field, A. Hayes, and R. Hess. Contour integration by the human visual system: evidence for a local "association field". *Vision Research*, 33(2):173–193, 1993. 29, 42

[92] G. Field and E. Chichilnisky. Information processing in the primate retina: circuitry and coding. *Annual Review of Neuroscience*, 30:1–30, 2007. 9

[93] D. Fortun, P. Bouthemy, and C. Kervrann. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding*, 134:1–21, 2015. 37

[94] C. Fowlkes, D. Martin, and J. Malik. Local figure-ground cues are valid for natural images. *J. of Vision*, 7(8):1–9, 2007. 31, 42

[95] Y. Fregnac and B. Bathelier. Cortical correlates of low-level perception: from neural circuits to percepts. *Neuron*, 88(7):110–126, 2015. 15, 18, 20

[96] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí. Yet another survey on image segmentation: Region and boundary information integration. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *European Conference on Computer Vision*, volume 2352, pages 408–422. Springer Berlin Heidelberg, 2002. 30

[97] P. Fries. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Science*, 9:474–480, 2005. 12, 13

[98] J. P. Frisby and J. V. Stone. *Seeing, Second Edition: The Computational Approach to Biological Vision.* The MIT Press, 2nd edition, 2010. 6, 15

[99] K. Fukushima. Neural network model for selective attention in visual pattern recognition and associative recall. *Applied Optics*, 26(23):4985–4992, 1987. 11

[100] F. Galasso, N. Nagaraja, T. Cardenas, T. Brox, and B. Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *IEEE International Conference on Computer Vision*, 2013. 44

[101] W. Geisler, J. Perry, B. Super, and D. Gallogly. Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41:711–724, 2001. 29, 42

[102] W. S. Geisler. Motion streaks provide a spatial code for motion direction. *Nature*, 400(6739):65–69, July 1999. 36

[103] W. Gerstner and W. Kistler. *Spiking Neuron Models*. Cambridge University Press, 2002. 46

[104] S. Gharaei, C. Talby, S. S. Solomon, and S. G. Solomon. Texture-dependent motion signals in primate middle temporal area. *Journal of Physiology*, 591:5671–5690, 2013. 35

[105] M. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements and actions. *Nature Reviews Neuroscience*, 4:179–192, 2003. 11, 21, 35

[106] M. Giese and G. Rizzolatti. Neural and computational mechanisms of action processing: interaction between visual and motor representations. *Neuron*, 88(1):167–180, 2015. 35

[107] A. Gilad, E. Meirovithz, and H. Slovin. Population responses to contour integration: early encoding of discrete elements and late perceptual grouping. *Neuron*, 2(389–402), 2013. 13, 35, 42

[108] C. D. Gilbert and W. Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013. 29

[109] M. Giuliani, X. Lagorce, F. Galluppi, and R. B. Benosman. Event-based computation of motion flow on a neuromorphic analog neural platform. *Frontiers in Neuroscience*, 10(9), Feb. 2016. 26

[110] T. Gollisch and M. Meister. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron*, 65(2):150–164, Jan. 2010. 9, 22, 23, 24

[111] G. H. Granlund. In search of a general picture processing operator. *Computer Graphics and Image Processing*, 8(2):155 – 173, 1978. 15

[112] S. Grossberg. How does the brain build a cognitive code? *Psychological Science*, 87(1):1–51, 1980. 40, 43

[113] S. Grossberg. A solution of the figure-ground problem for biological vision. *Neural Networks*, 6:463–483, 1993. 32, 42

[114] S. Grossberg and E. Mingolla. Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychological review*, 92(2):173–211, 1985. 29, 31, 32, 42

[115] S. Grossberg, E. Mingolla, and C. Pack. A neural model of motion processing and visual navigation by cortical area MST. *Cerebral Cortex*, 9(8):878–895, Dec. 1999. 35

[116] S. Grossberg, E. Mingolla, and W. D. Ross. Visual brain and visual perception: how does the cortex do perceptual grouping? *Trends in Neurosciences*, 20(3):106–111, 1997. 31, 42

[117] S. Grossberg, E. Mingolla, and L. Viswanathan. Neural dynamics of motion integration and segmentation within and across apertures. *Vision Research*, 41(19):2521–2553, 2001. 19, 41, 42

[118] A. Grunewald, D. Bradley, and R. Andersen. Neural correlates of structure-from-motion perception in macaque area V1 and MT. *Journal of Neuroscience*, 22(14):6195–6207, 2002. 36

[119] U. Güçlü and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, 35(27):10005–10014, July 2015. 43

[120] M. Gur. Space reconstruction by primary visual cortex activity: a parallel, non-computational mechanism of object representation. *Trends in Neurosciences*, 38(4):207–216, 2015. 6, 11, 36

[121] B. Hassenstein and R. W. Systemtheoretische analyse der zeit, reihenfolgen und vorzeichenauswertung. *In The Bewegungsperzeption Des weevil Chlorophanus. Z. Naturforsch.*, 1956. 34, 37

[122] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188 – 194, 2005. 26, 32, 42

[123] M. Hayhoe, J. Lachter, and J. Feldman. Integration of form across saccadic eye movements. *Perception*, 20:392–402, 1991. 33

[124] J. H. Hedges, Y. Gartshteyn, A. Kohn, N. C. Rust, M. N. Shadlen, W. T. Newsome, and J. A. Movshon. Dissociation of neuronal and psychophysical responses to local and global motion. *Current Biology*, 21(23):2023–2028, 2011. 10

[125] D. Heeger. Optical flow using spatiotemporal filters. *The International Journal of Computer Vision*, 1(4):279–302, Jan. 1988. 38, 41, 42

[126] J. Hegdé and D. Felleman. Reappraising the Functional Implications of the Primate Visual Anatomical Hierarchy. *The Neuroscientist*, 13(5):416–421, Oct. 2007. 10

[127] J. Hegdé and D. C. Van Essen. A comparative study of shape representation in macaque visual areas V2 and V4. *Cerebral Cortex*, 17(5):1100–1116, 2007. 10

[128] M. Helmstaedter, K. Briggman, S. Turaga, V. Jain, S. Seung, and W. Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500:168–174, 2013. 5

[129] J. Hérault. *Vision: Images, Signals and Neural Networks: Models of Neural Processing in Visual Perception*. World Scientific, 2010. 6, 15, 24, 26, 42

[130] J. Hérault and B. Durette. Modeling visual perception for image processing. In F. Sandoval, A. Prieto, J. Cabestany, and M. Grana, editors, *Computational and Ambient Intelligence : 9th International Work-Conference on Artificial Neural Networks, IWANN 2007*, 2007. 42

[131] D. Heslip, T. Ledgeway, and P. McGraw. The orientation tuning of motion streak mechanisms revealed by masking. *Journal of Vision*, 13(9):376, 2013. 36

[132] C. Hilario Gomez, K. Medathati, P. Kornprobst, V. Murino, and D. Sona. Improving freak descriptor for image classification. In *ICVS*, 2015. 25

[133] E. C. Hildreth and C. Koch. The analysis of visual motion: From computational theory to neuronal mechanisms. *Annual Review of Neuroscience*, 10(1):477–533, 1987. PMID: 3551763. 15, 35

[134] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006. 31

[135] G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006, 2006. 8

[136] S. Hochstein and M. Ahissar. View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*, 5(791-804), 36. 11

[137] D. Hoiem, A. A. Efros, and M. Hebert. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, 91(3):328–346, Feb. 2011. 32, 42

[138] S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. In N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2015. 30

[139] S. Hong, J. Oh, B. Han, and H. Lee. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *International Conference on Computer Vision and Pattern Recognition*, 2016. 30, 31

[140] B. Horn and B. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17:185–203, 1981. 37

[141] X. Huang, T. Albright, and G. Stoner. Adaptive surround modulation in cortical area MT. *Neuron*, 53:761–770, 2007. 36, 42

[142] A. C. Huk. Multiplexing in the primate motion pathway. *Vision Research*, 62(0):173 – 180, 2012. 42

[143] J. Hupé, A. James, B. Payne, S. Lomber, P. Girard, and J. Bullier. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*, 394:784–791, 1998. 36, 42

[144] G. Ibos and D. Freedman. Dynamic integration of task-relevant visual features in posterior parietal cortex. *Neuron*, 83(6):1468–80, 2014. 12

[145] J. Issacson and M. Scanziani. How inhibition shapes cortical activity. *Neuron*, 72:231–240, 2011. 12

[146] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001. 12

[147] D. Jancke, F. Chavane, S. Naaman, and A. Grinvald. Imaging cortical correlates of illusion in early visual cortex. *Nature*, 428:423–426, 2004. 13

[148] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *IEEE International Conference on Computer Vision*, pages 2146–2153, 2009. 40

[149] D. Jeurissen, M. Self, and P. Roelfsema. Serial grouping of 2d-image regions with object-based attention, 2016. under revision. 32, 33

[150] D. Jeurissen, M. W. Self, and P. R. Roelfsema. Surface reconstruction, figure-ground modulation, and border-ownership. *Cognitive Neuroscience*, 4(1):50–52, 2013. PMID: 24073702. 42

[151] P. Jolicoeur, S. Ullman, and M. Mackay. Curve tracing: A possible basic operation the perception of spatial relations. *Memory and Cognition*, 14(2):129–140, 1986. 30, 42

[152] H. E. Jones, I. M. Andolina, B. Ahmed, S. D. Shipp, J. T. C. Clements, K. L. Grieve, J. Cudeiro, T. E. Salt, and A. M. Sillito. Differential feedback modulation of center and surround mechanisms in parvocellular cells in the visual thalamus. *The Journal of Neuroscience*, 32(45):15946–15951, 2012. 9

[153] H. Kafaligonul, B. Breitmeyer, and H. Ogmen. Feedforward and feedback processes in vision. *Frontiers in Psychology*, 6(279), 2015. 12

[154] M. Kapadia, M. Ito, C. Gilbert, and G. Westheimer. Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. *Neuron*, 50:35–41, 1995. 29

[155] M. K. Kapadia, G. Westheimer, and C. D. Gilbert. Spatial distribution of contextual interactions in primary visual cortex and in visual perception. *Journal of Neurophysiology*, 84(4):2048–2062, 2000. 29, 42

[156] D. B. Kastner and S. A. Baccus. Insights from the retina into the diverse and general computations of adaptation, detection, and prediction. *Current Opinion in Neurobiology*, 25:63–69, Apr. 2014. 9, 22, 42

[157] P. Kellman and T. Shipley. A theory of visual interpolation in object perception. *Cognitive Psychology*, 23:141–221, 1991. 31, 42

[158] P. Khorsand, T. Moore, and A. Soltani. Combined contributions of feedforward and feedback inputs to bottom-up attention. *Frontiers in Psychology*, 6(155), 2015. 12

[159] H. Kim, D. Angelaki, and G. DeAngelis. A novel role for visual perspective cues in the neural computation of depth. *Nature Neuroscience*, 18(1):129–137, 2015. 36

[160] J. S. Kim, M. J. Greene, A. Zlateski, K. Lee, M. Richardson, S. C. Turaga, M. Purcaro, M. Balkam, A. Robinson, B. F. Behabadi, M. Campos, W. Denk, H. S. Seung, and Eye-Wirers. Space-time wiring specificity supports direction selectivity in the retina. *Nature*, 509(331-336), 2014. 5, 24

[161] J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984. 27

[162] J. Koenderink, W. Richards, and A. van Doorn. Blow-up: a free lunch? *i-Perception*, 3(2):141–145, 2012. 27

[163] K. Koffka. *Principles of Gestalt psychology*. Routledge & Kegan Paul Ltd., London, 1935. 29

[164] N. Kogo and J. Wagemans. The "side" matters: How configurality is reflected in completion. *Cognitive Neuroscience*, 4(1):31–45, 2013. PMID: 24073697. 27, 42

[165] P. Kornprobst and G. Médioni. Tracking segmented objects using tensor voting. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 118–125, Hilton Head Island, South Carolina, June 2000. IEEE Computer Society. 31

[166] N. Kriegeskorte. Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, 3(3):363–373, 2009. 43

[167] N. Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446, Nov. 2015. 18

[168] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In P. Bartlett, F. Pereira, C. Burges, L. Botton, and K. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, 2012. 30

[169] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1847–1871, Aug. 2013. 6, 7, 17, 40

[170] J. Kubilius, J. Wagemans, and H. Op de Beeck. A conceptual framework of computations in mid-level vision. *Frontiers in Computational Neuroscience*, 8:158:1–19, 2014. 6

[171] J. Kung, H. Yamaguchi, C. Liu, G. Johnson, and M. Fairchild. Evaluating HDR rendering algorithms. *ACM Transactions on Applied Perception*, 4(2), July 2007. 25

[172] V. Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *The Journal of Neuroscience*, 15(2):1605–1615, 1995. 29, 31, 42

[173] V. Lamme, V. Rodriguez-Rodriguez, and H. Spekreijse. Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cerebral Cortex*, 9:406–413, 1999. 29

[174] V. Lamme, K. Zipser, and H. Spekreijse. Figure-ground activity in primary visual cortex is suppressed by anesthesia. *PNAS*, 95:3263–3268, 1998. 31, 42

[175] V. A. F. Lamme and P. R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579, 2000. 10, 12, 30, 35

[176] M. Lappe. Functional consequences of an integration of motion and stereopsis in area MT of monkey extrastriate visual cortex. *Neural Comput.*, 8(7):1449–1461, 1996. 36, 42

[177] O. Layton and N. Browning. A unified model of heading and path perception in primate MSTd. *PLoS Computational Biology*, 10(2):e1003476, 2014. 35

[178] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015. 20

[179] T. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of Optical Society of America A*, 20(7), 2003. 11, 12

[180] T. Lee and M. Nguyen. Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences*, 98(4):1907, 2001. 10

[181] S. Lehky, M. Sereno, and A. Sereno. Population coding and the labeling problem: extrinsic versus intrinsic representations. *Neural Computation*, 25(9):2235–2264, 2013. 14

[182] W. Li, V. Piech, and C. Gilbert. Learning to link visual contours. *Neuron*, 57:442–451, 2008. 29, 42

[183] Z. Li. The immersed interface method using a finite element formulation. *Applied Numerical Mathematics*, 27(3):253–267, 1998. 31

[184] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 26, 27, 42

[185] T. Lindeberg. Feature detection with automatic scale selection. *The International Journal of Computer Vision*, 30(2):77–116, 1998. 30

[186] S. Lisberger. Visual guidance of smooth-pursuit eye movements: sensation, action and what happens in between. *Neuron*, 66(4):477–491, 2010. 35

[187] D. Liu, J. Gu, Y. Hitomi, M. Gupta, T. Mitsunaga, and S. Nayar. Efficient space-time sampling with pixel-wise coded exposure for high-speed imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):248–260, 2014. 25

[188] S.-C. Liu and T. Delbruck. Neuromorphic sensory systems. *Current Opinion in Neurobiology*, 20:1–8, 2010. 26, 42

[189] S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, and R. Douglas, editors. *Event-Based Neuromorphic Systems*. Wiley, Jan. 2015. 6, 26

[190] Y. S. Liu, C. F. Stevens, and T. O. Sharpee. Predictable irregularities in retinal receptive fields. *Proceedings of the National Academy of Sciences*, 106(38):16499–16504, Sept. 2009. 26

[191] M. Livingstone and D. H. Hubel. Segregation of form, color, movement and depth: anatomy, physiology and perception. *Science*, 240(740-749), 1988. 11

[192] H. Lorach, R. Benosman, O. Marre, S.-H. Ieng, J. A. Sahel, and S. Picaud. Artificial retina: the multichannel processing of the mammalian retina achieved with a neuromorphic asynchronous light acquisition device. *Journal of Neural Engineering*, 9(6):066004, Oct. 2012. 24, 26, 41, 42, 46

[193] D. G. Lowe. Local feature view clustering for 3d object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 682–688, 2001. 26

[194] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981. 37

[195] L. Luo, E. Callaway, and K. Svoboda. Genetic dissection of neural circuits. *Neuron*, 57(5):634–660, 2008. 12

[196] S. Lyu and E. P. Simoncelli. Nonlinear extraction of independent components of natural images using radial gaussianization. *Neural Comput.*, 21(6):1485–1519, June 2009. 40

[197] O. Mac Aodha, A. Humayun, M. Pollefeys, and G. Brostow. Learning a confidence measure for optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1107–1120, 2013. 38

[198] V. Mante and M. Carandini. Mapping of stimulus energy in primary visual cortex. *Journal of Neurophysiology*, 94:788—798, Mar. 2005. 35, 42

[199] R. Marc, B. Jones, C. Watt, J. Anderson, C. Sigulinsky, and S. Lauritzen. Retinal connectomics: towards complete, accurate networks. *Progress in Retinal and Eye Research*, 37:141–162, 2013. 22

[200] N. Markov, M. Ercsey-Ravasz, D. Van Essen, K. Knoblauch, Z. Toroczkai, and H. Kennedy. Cortical high-density counterstream architectures. *Science*, 342, 2013. 5, 7, 10

[201] N. T. Markov, J. Vezoli, P. Chameau, A. Falchier, R. Quilodran, C. Huissoud, C. Lamy, P. Misery, P. Giroud, S. Ullman, P. Barone, C. Dehay, K. Knoblauch, and H. Kennedy. Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, 522(1):225–259, 2014. 17

[202] D. Marr. *Vision*. W.H. Freeman and Co., 1982. 5, 7, 15

[203] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society London, B*, 207:187–217, 1980. 30

[204] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE International Conference on Computer Vision*, volume 2, pages 416–423, 2001. 26, 42, 44

[205] A. Martinez-Alvarez, A. Olmedo-Payá, S. Cuenca-Asensi, J. M. Ferrandez, and E. Fernandez. RetinaStudio: A bioinspired framework to encode visual information. *Neurocomputing*, 114:45–53, Aug. 2013. 41

[206] R. H. Masland. Cell populations of the retina: The proctor lecture. *Investigative Ophthalmology and Visual Science*, 52(7):4581–4591, June 2011. 22, 23, 42

[207] R. H. Masland. The Neuronal Organization of the Retina. *Neuron*, 76(2):266–280, Oct. 2012. 22, 23, 42

[208] T. Masquelier and S. Thorpe. Learning to recognize objects using waves of spikes and spike timing-dependent plasticity. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, page 1–8, 2010. 46

[209] G. Masson and L. Perrinet. The behavioral receptive field underlying motion integration for primate tracking eye movements. *Neurosciences and BioBehavioral Reviews*, 36(1):1–25, 2012. 35

[210] G. Mather, A. Pavan, R. M. Bellacosa, and C. Casco. Psychophysical evidence for interactions between visual motion and form processing at the level of motion integrating receptive fields. *Neuropsychologia*, 50(1):153 – 159, 2012. 11, 36

[211] J. Maunsell and D. Van Essen. Functional properties of neurons in middle temporal visual area of the macaque monkey. ii. binocular interactions and sensitivity to binocular disparity. *Journal of Neurophysiology*, 49:1148–1167, 1983. 42

[212] J. McCarthy, D. Cordeiro, and G. Caplovitz. Local form–motion interactions influence global form perception. *Attention, Perception, & Psychophysics*, 74(5):816–823, 2012. 36

[213] J. S. McDonald, C. W. Clifford, S. S. Solomon, S. C. Chen, and S. G. Solomon. Integration and segregation of multiple motion signals by neurons in area MT of primate. *J Neurophysiol.*, 2014. 38

[214] P. W. McOwan and A. Johnston. Motion transparency arises from perceptual grouping: evidence from luminance and contrast modulation motion displays. *Current Biology*, 6(10):1343 – 1346, 1996. 33

[215] N. V. K. Medathati, M. Chessa, G. S. Masson, P. Kornprobst, and F. Solari. Adaptive motion pooling and diffusion for optical flow. Technical Report 8695, INRIA, Mar. 2015. 38, 39

[216] G. Medioni, M. Lee, and C. Tang. *A Computational Framework for Segmentation and Grouping*. Elsevier, 2000. 31

[217] L. Merabet, A. Desautels, K. Minville, and C. Casanova. Motion integration in a thalamic visual nucleus. *Nature*, 396(265–268), 1998. 9

[218] P. Merolla, J. Arthur, R. Alvarez-Icaza, A. Cassidy, J. Sawada, F. Akopyan, B. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. Esser, R. Appuswamy, B. Taba, A. Amir, M. Flickner, W. Risk, R. Manohar, and D. Modha. Artificial brains. a million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(668-673), 2014. 5

[219] L. Meylan, D. Alleysson, and S. Süsstrunk. Model of retinal local adaptation for the tone mapping of color filter array images. *Journal of Optical Society of America. A*, 24(9):2807–2816, 2007. 25, 46

[220] A. D. Milner and M. A. Goodale. Two visual systems re–viewed. *Neuropsychologia*, 46, 2008. 7

[221] P. Milner. A model for visual shape recognition. *Psychological Review*, 81(6):521–535, 1974. 11

[222] P. Mineault, F. Khawaja, D. Butts, and C. Pack. Hierarchical processing of complex motion along the primate dorsal visual pathways. *Proceedings of the National Academy of Sciences*, 109(972-980), 2012. 36

[223] A. Mishra and Y. Aloimonos. Active segmentation. *International Journal of Humanoid Robotics (IJHR)*, 6(3):361—386, 2009. 31

[224] A. Mishra, Y. Aloimonos, L.-F. Cheong, and A. Kassim. Active visual segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):639–653, 2012. 31

[225] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovsky, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, Feb. 2015. 20

[226] B. Motter. Focal attention produces spatially selective processing in visual cortical areas v1, v2 and v4 in the presence of competing stimuli. *Journal of Neurophysiology*, 70(3):909–919, 1993. 12

[227] K. Muchungi and M. Casey. Simulating Light Adaptation in the Retina with Rod-Cone Coupling. In *ICANN*, pages 339–346. Springer Berlin Heidelberg, Berlin, Heidelberg, Sept. 2012. 25

[228] L. Muller, A. Reynaud, F. Chavane, and A. Destexhe. The stimulus-evoked population response in visual cortex of awake monkey is a propagating wave. *Nature Communications*, 5(3675), 2014. 13

[229] D. Mumford. On the computational architecture of the neocortex. I. the role of the thalamo-cortical loop. *Biological Cybernetics*, 65:135–145, 1991. 9

[230] J. Nadler, M. Nawrot, D. Angelaki, and G. DeAngelis. MT neurons combine visual motion with a smooth eye movement signal to code depth-sign from motion parallax. *Neuron*, 63(4):523–532, 2009. 36, 42

[231] K. Nakayama. Biological image motion processing: A review. *Vision Research*, 25(5):625 – 660, 1985. 35, 42

[232] A. Nandy, T. Sharpee, J. Reynolds, and J. Mitchell. The fine structure of shape tuning in area V4. *Neuron*, 78(1102-1115), 2013. 8

[233] H. Nasser, S. Kraria, and B. Cessac. Enas: a new software for neural population analysis in large scale spiking networks. In *Springer Series in Computational Neuroscience*. Organization for Computational Neurosciences, July 2013. 24

[234] J. Nassi, C. Gomez-Laberge, G. Kreiman, and R. Born. Corticocortical feedback increases the spatial extent of normalization. *Frontiers in Systems Neurosciences*, 8(105), 2014. 36

[235] H. Neumann and E. Mingolla. Computational neural models of spatial integration in perceptual grouping. In T. F. Shipley and P. J. Kellman, editors, *From Fragments to Objects: Grouping and Segmentation in Vision*, pages 353–400. Amsterdam: Elsevier, 2001. 31, 42

[236] Z. Ni, C. Pacoret, R. Benosman, S. Ieng, and S. Régnier. Asynchronous event-based high speed vision for microparticle tracking. *Journal of Microscopy*, 245(3):236–244, Nov. 2011. 26

[237] S. Nishimoto and J. L. Gallant. A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. *The Journal of Neuroscience*, 31(41):14551–14564, 2011. 35, 37, 42

[238] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision*, pages 1520–1528, Santiago, Chile, Dec. 2015. 30

[239] H. Nothdurft. Texture segmentation and pop-out from orientation contrast. *Vision Research*, 31(6):1073–1078, 1991. 27

[240] S. Nowlan and T. Sejnowski. Filter selection model for motion segmentation and velocity integration. *J. Opt. Soc. Am. A*, 11(12):3177–3199, 1994. 19, 37, 38, 41, 42

[241] B. Odermatt, A. Nikolaev, and L. Lagnado. Encoding of luminance and contrast by linear and nonlinear synapses in the retina. *Neuron*, 73(4):758 – 773, 2012. 24

[242] P. O'Herron and R. von der Heydt. Representation of object continuity in the visual cortex. *J. of Vision*, 11(2):12, 1–9, 2011. 29

[243] T. Ohshiro, D. E. Angelaki, and G. C. DeAngelis. A normalization model of multisensory integration. *Nature Neuroscience*, 14(6):775–782, May 2011. 36, 42

[244] G. A. Orban. Higher order visual processing in macaque extrastriate cortex. *Physiological Reviews*, 88(1):59–89, 2008. 7, 11, 17, 34, 35

[245] G. Orchard, C. Meyer, R. Etienne-Cummings, C. Posch, N. Thakor, and R. Benosman. Hfirst: A temporal approach to object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10), Oct. 2015. 26

[246] C. Pack and R. Born. Cortical mechanisms for the integration of visual motion. In R. H. Masland, T. D. Albright, T. D. Albright, R. H. Masland, P. Dallos, D. Oertel, S. Firestein, G. K. Beauchamp, M. C. Bushnell, A. I. Basbaum, J. H. Kaas, and E. P. Gardner, editors, *The Senses: A Comprehensive Reference*, pages 189 – 218. Academic Press, New York, 2008. 34, 42

[247] N. Pal and S. Pal. A review of image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993. 26, 30

[248] C. Pandarinath, J. Victor, and S. Nirenberg. Symmetry Breakdown in the ON and OFF Pathways of the Retina at Night: Functional Implications. *The Journal of neuroscience*, 30(30):10006–10014, 2010. 26

[249] P. Parent and S. Zucker. Trace inference, curvature consistency, and curve detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):823–839, 1989. 31, 41

[250] L. Perrinet. *Biologically inspired computer vision*, chapter Sparse models for computer vision, pages 319–346. Number 14. Wiley, 2015. 13

[251] L. Perrinet, M. Samuelides, and S. Thorpe. Coding static natural images using spiking event times: do neuron cooperate? *IEEE Transactions in Neural Networks and Learning Systems*, 15(5):1164–1175, 2004. 14

[252] L. U. Perrinet and G. S. Masson. Motion-based prediction is sufficient to solve the aperture problem. *Neural Computation*, 24(10):2726–2750, 2012. 12, 40

[253] J. A. Perrone. A neural-based code for computing image velocity from small sets of middle temporal (MT/V5) neuron inputs. *Journal of Vision*, 12(8), 2012. 19, 35, 41, 42

[254] L. Pessoa, E. Thompson, and A. Noë. Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and brain sciences*, 21:723–802, 1998. 30

[255] E. Peterhans and R. Von der Heydt. Subjective contours: bridging the gap between psychophysics and physiology. *Trends in Neurosciences*, 14(3):112–119, 1991. 10, 42

[256] M. A. Peterson and E. Salvagio. Inhibitory competition in figure-ground perception: Context and convexity. *Journal of Vision*, 8(16), 2008. 27, 42

[257] J. Petitot. An introduction to the Mumford-Shah segmentation model. *Journal of Physiology - Paris*, 97:335–342, 2003. 32

[258] M. Petrou and A. Bharat. *Next generation artificial vision systems: Reverse engineering the human visual system.* Artech House Series Bioinformatics & Biomedical Imaging, 2008. 6, 15

[259] V. Piech, W. Li, G. Reeke, and C. Gilbert. Network model of top-down influences on local gain and contextural interactions in visual cortex. *Proceedings of the National Academy of Sciences - USA*, 110(43):4108–4117, 2013. 10

[260] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008. 23, 42

[261] N. Pinto and D. Cox. *GPU Meta-Programming: A Case Study in Biologically-Inspired Machine Vision. GPU Computing Gems*, volume 2, chapter 33. Elsevier, 2012. 5

[262] N. Pinto, D. Doukhan, J. DiCarlo, and D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, 5(11), 2009. 19, 43

[263] H. Plesser, J. Eppler, A. Morrison, M. Diesmann, and M.-O. Gewaltig. Efficient parallel simulation of large-scale neuronal networks on clusters of multiprocessor computers. In A.-M. Kermarrec, L. Bougé, and T. Priol, editors, *Euro-Par 2007 Parallel Processing*, volume 4641 of *Lecture Notes in Computer Science*, pages 672–681. Springer Berlin Heidelberg, 2007. 5

[264] T. Poggio. The levels of understanding framework, revised. *Perception*, 41(9):1017–1023, 2012. 15

[265] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317(6035):314–319, Sept. 1985. 28

[266] M. Pomplun and J. Suzuki, editors. *Developing and Applying Biologically-Inspired Vision Systems: Interdisciplinary Concepts*. IGI Global, 2012. 6

[267] J. Poort, F. Raudies, A. Wannig, V. Lamme, N. H., and P. Roelfsema. The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron*, 108(5):1392–1402, 2012. 42

[268] G. Portelli, J. Barrett, E. Sernagor, T. Masquelier, and P. Kornprobst. The wave of first spikes provides robust spatial cues for retinal information processing. Research Report RR-8559, INRIA, July 2014. 26

[269] C. Posch, D. Matolin, and R. Wohlgenannt. A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2011. 26, 42

[270] T. Potjans and D. M. The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model. *Cerebral Cortex*, 24(785-806), 2014. 5

[271] A. Pouget, J. Beck, W. Ma, and P. Latham. Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178, 2013. 13, 14

[272] A. Pouget, P. Dayan, and R. Zemel. Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132, 2000. 43

[273] A. Pouget, P. Dayan, and R. Zemel. Inference and computation with population codes. *Annual Review of Neuroscience*, 26:381–410, 26. 14

[274] N. Priebe, C. Cassanello, and S. Lisberger. The neural representation of speed in macaque area MT/V5. *Journal of Neuroscience*, 23(13):5650–5661, July 2003. 35

[275] N. Priebe, S. Lisberger, and A. Movshon. Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex. *The Journal of Neuroscience*, 26(11):2941—2950, 2006. 10, 42

[276] Z. Pylyshyn. Is vision continuous with cognition? the case for cognitive impenetrability of visual perception. *Behavioral and brain sciences*, 22(3):341–365, 1999. 19

[277] N. Qian and R. A. Andersen. A physiological model for motion-stereo integration and a unified explanation of Pulfrich-like phenomena. *Vision Research*, 37:1683–1698, 1997. 36, 42

[278] J. Rankin, I. Meso, Andrew, S. Masson, Guillaume, O. Faugeras, and P. Kornprobst. Bifurcation study of a neural fields competition model with an application to perceptual switching in motion integration. *Journal of Computational Neuroscience*, 36(2):193–213, 2014. 46

[279] R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*, 2(1):79–87, 1999. 40

[280] M. J. Rasch, M. Chen, S. Wu, H. D. Lu, and A. W. Roe. Quantitative inference of population response properties across eccentricity from motion-induced maps in macaque V1. *Journal of Neurophysiology*, 109(5):1233–1249, 2013. 38

[281] F. Raudies, E. Mingolla, and H. Neumann. A model of motion transparency processing with local center-surround interactions and feedback. *Neural Computation*, 23:2868–2914, 2011. 37

[282] F. Raudies, E. Mingolla, and H. Neumann. Active gaze control improves optic flow-based segmentation and steering. *PLoS ONE*, 7(6):1–19, 06 2012. 33

[283] F. Raudies and H. Neumann. A neural model of the temporal dynamics of figure–ground segregation in motion perception. *Neural Networks*, 23(2):160 – 176, 2010. 42

[284] X. Ren, C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision – ECCV 2006*, volume 3952 of *Lecture Notes in Computer Science*, pages 614–627. Springer Berlin Heidelberg, 2006. 32, 41, 42

[285] A. Reynaud, G. Masson, and F. Chavane. Dynamics of local input normalization result from balanced short- and long-range intracortical interactions in area V1. *Journal of Neuroscience*, 32:12558–12569, 2012. 12, 13, 35

[286] J. Reynolds and D. Heeger. The normalisation model of attention. *Neuron*, 61:168–185, 2009. 12, 16

[287] J. Reynolds, T. Pasternak, and R. Desimone. Attention increases sensitivity of v4 neurons. *Neuron*, 26:703–714, 2000. 12

[288] H. Rodman and T. Albright. Coding of visual stimulus velocity in area MT of the macaque. *Vision Research*, 27(12):2035–2048, 1987. 10

[289] A. J. Rodriguez Sanchez and J. K. Tsotsos. The roles of endstopped and curvature tuned computations in a hierarchical representation of 2d shape. *PLoS ONE*, 7(8):e42058, 2012. 32, 41, 42

[290] P. Roelfsema. Elemental operations in vision. *Trends in Cognitive Sciences*, 9(5):226–233, 2005. 11, 30

[291] P. R. Roelfsema. Cortical algorithms for perceptual grouping. *Annual Review of Neuroscience*, 29(1):203–227, 2006. PMID: 16776584. 29

[292] P. R. Roelfsema and R. Houtkamp. Incremental grouping of image elements in vision. *Attention, Perception, & Psychophysics*, 73(8):2542–2572, 2011. 29, 30

[293] P. R. Roelfsema, V. A. F. Lamme, and H. Spekreijse. The implementation of visual routines. *Vision Research*, 40(10–12):1385–1411, 2000. 7, 12

[294] P. R. Roelfsema, V. A. F. Lamme, H. Spekreijse, and H. Bosch. Figure/ground segregation in a recurrent network architecture. *J. of Cognitive Neuroscience*, 14(4):525–537, 2002. 12, 29

[295] P. R. Roelfsema, M. Tolboom, and P. S. Khayat. Different processing phases for features, figures, and selective attention in the primary visual cortex. *Neuron*, 56(5):785 – 792, 2007. 29

[296] P. Rogister, R. Benosman, and S. H. Ieng. Asynchronous event-based binocular stereo matching. *IEEE Transactions in Neural Networks and Learning Systems*, pages 347–353, 2012. 26

[297] E. Rolls and G. Deco. *The noisy brain: stochastic dynamics as a principle of brain function.* Oxford university press, 2010. 14

[298] G. Rousselet, S. Thorpe, and M. Fabre-Thorpe. How parallel is visual processing in the ventral path? *TRENDS in Cognitive Sciences*, 8(8):363–370, Aug. 2004. 11, 12

[299] M. Rucci and J. Victor. The unsteady eye: an information-processing stage, not a bug. *Trends in Neurosciences*, 38(4):195–206, 2015. 9

[300] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008. 44

[301] N. Rust, V. Mante, E. Simoncelli, and J. Movshon. How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, 9:1421–1431, 2006. 19, 35, 36, 37, 38, 42

[302] N. Rust, O. Schwartz, J. Movshon, and E. Simoncelli. Spatiotemporal elements of macaque V1 receptive fields. *Neuron*, 46:945–956, 2005. 42

[303] T. Sato, I. Nauhaus, and M. Carandini. Traveling waves in visual cortex. *Neuron*, 75:218–229, 2012. 13

[304] W. Scheirer, S. Anthony, K. Nakayama, and D. Cox. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1679–1686, 2014. 44

[305] O. Scherzer and J. Weickert. Relations between regularization and diffusion filtering. *Journal of Mathematical Imaging and Vision*, 12(1):43–63, Feb. 2000. 38

[306] H. Scholte, J. Jolij, J. Fahrenfort, and V. Lamme. Feedforward and recurrent processing in scene segmentation: electroencephalography and functional magnetic resonance imaging. *J. of Cognitive Neuroscience*, 20(11):2097–2109, 2008. 29, 32

[307] M. W. Self, T. van Kerkoerle, H. Super, and P. R. Roelfsema. Distinct roles of the cortical layers of area V1 in figure-ground segregation. *Current Biology*, 23(21):2121 – 2129, 2013. 36, 42

[308] A. Sellent, M. Eisemann, B. Goldlucke, D. Cremers, and M. Magnor. Motion field estimation from alternate exposure images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1577–1589, 2011. 38

[309] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences (PNAS)*, 104(15):6424–6429, 2007. 8

[310] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, Mar. 2007. 8

[311] M. Shadlen and J. Movshon. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron*, 24(1):67–77, 1999. 13

[312] M. Shamir. Emerging principles of population coding: in search for the neural code. *Current Opinion in Neurobiology*, 25:140–148, 2014. 13

[313] R. Shapley. Visual sensitivity and parallel retinocortical channels. *Annual Review of Psychology*, 41:635–658, 1990. 23

[314] R. Shapley and C. Enroth-Cugell. Visual adaptation and retinal gain controls. *Progress in retinal research*, 3:263–346, 1984. 22, 42

[315] R. Shapley, M. Hawken, and D. Ringach. Dynamics of orientation selectivity in the primate visual cortex and the importance of cortical inhibition. *Neuron*, 38(5):689–699, 2003. 14

[316] G. Sheperd and S. Grillner, editors. *Handbook of brain microcircuits*. Oxford University Press, 2010. 15

[317] J. S. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 30

[318] A. Sigman, A. Cecchi, C. Gilbert, and M. Magnaso. On a common circle: Natural scenes and gestalt rules. *PNAS*, 98(4):1935–1940, 2001. 31, 42

[319] M. Silies, D. Gohl, and T. Clandinin. Motion-detecting circuits in flies: Coming into view. *Annual Review of Neuroscience*, 37(1):307–327, 2014. PMID: 25032498. 34

[320] E. Simoncelli and D. Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 38:743–761, 1998. 35, 36, 38, 39, 42

[321] E. Simoncelli and B. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001. 43

[322] C. Simoncini, L. Perrinet, A. Montagnini, P. Mamassian, and G. Masson. More is not always: adaptive gain control explains dissociation between perception and action. *Nature Neuroscience*, 15(11):1586–1603, 2012. 36, 37

[323] W. Singer. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24(1):49–65, 1999. 13

[324] A. Smolyanskaya, D. A. Ruff, and R. T. Born. Joint tuning for direction of motion and binocular disparity in macaque MT is largely separable. *Journal of Neurophysiology*, 2013. 36, 42

[325] R. J. Snowden, S. Treue, R. G. Erickson, and R. A. Andersen. The response of area MT and V1 neurons to transparent motion. *The Journal of Neuroscience*, 11(9):2768–2785, 1991. 36

[326] F. Solari, M. Chessa, K. Medathati, and P. Kornprobst. What can we expect from a V1-MT feedforward architecture for optical flow estimation? *Signal Processing: Image Communication*, 2015. 19, 38, 39, 41, 42, 46

[327] R. Squire, B. Noudoost, R. Schafer, and T. Moore. Prefrontal contributions to visual selective attention. *Annual Review of Neuroscience*, 36:451–466, 2013. 12

[328] A. N. Stein and M. Hebert. Occlusion boundaries from motion: Low-level detection and mid-level reasoning. *International Journal of Computer Vision*, 82(3):325–357, 2009. 31

[329] G. R. Stoner and T. D. Albright. Neural correlates of perceptual motion coherence. *Nature*, 358:412–414, 1992. 36

[330] Y. Sugita. Grouping of image fragments in primary visual cortex. *Nature*, 401(6750):269–272, 1999. 10

[331] D. Sun, S. Roth, T. Darmstadt, and M. Black. Secrets of optical flow estimation and their principles. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2010. 37

[332] P. Sundberg, T. Brox, M. Maire, P. Arbeláez, and J. Malik. Occlusion boundary detection and figure/ground assignment from optical flow. In *International Conference on Computer Vision and Pattern Recognition*, 2011. 31, 32

[333] O. Temam and R. Héliot. Implementation of signal processing tasks on neuromorphic hardware. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1120 –1125, 2011. 5

[334] A. Thiele, K. Dobkins, and T. Albright. Neural correlates of chromatic motion perception. *Neuron*, 32(351-358), 2001. 11, 17

[335] A. Thiele and J. Perrone. Speed skills: measuring the visual speed analyzing properties of primate MT neurons. *Nature Neuroscience*, 4(5):526–532, 2001. 42

[336] W. Thoreson and S. Mangel. Lateral interactions in the outer retina. *Prog Retin Eye Res.*, 31(5):407–441, 2012. 22, 42

[337] S. Thorpe. The speed of categorization in the human visual system. *Neuron*, 62(2):168–170, 2009. 11, 12

[338] S. Thorpe, A. Delorme, and R. Van Rullen. Spike-based strategies for rapid processing. *Neural Networks*, 14(6-7):715–725, 2001. 14

[339] E. Tlapale, P. Kornprobst, G. S. Masson, and O. Faugeras. A neural field model for motion estimation. In S. Verlag, editor, *Mathematical Image Processing*, volume 5 of *Springer Proceedings in Mathematics*, pages 159–180, 2011. 18, 38, 46

[340] E. Tlapale, G. S. Masson, and P. Kornprobst. Modelling the dynamics of motion integration with a new luminance-gated diffusion mechanism. *Vision Research*, 50(17):1676–1692, Aug. 2010. 12, 17, 19, 36, 37, 40, 41, 42, 46

[341] S. Tschechne and H. Neumann. Hierarchical representation of shapes in visual cortex - from localized features to figural shape segregation. *Frontiers in Computational Neuroscience*, 8(93), 2014. 32, 41, 42

[342] S. Tschechne, R. Sailer, and H. Neumann. Bio-inspired optic flow from event-based neuromorphic sensor input. *Artificial Neural Networks in Pattern Recognition*, 8774:171–182, 2014. 26, 41

[343] J. Tsotsos. *Spatial vision in humans and robots*, chapter An inhibitory beam for attentional selection, pages 313–331. Cambridge University Press, 1993. 11

[344] J. Tsotsos. *A computational perspective on visual attention*. MIT Press, 2011. 12, 19, 21

[345] J. Tsotsos, S. Culhane, W. Kei Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78:507–545, 1995. 11, 12

[346] J. K. Tsotsos. It's all about the constraints. *Current Biology*, 24(18):854–858, Sept. 2014. 6, 20, 43

[347] J. K. Tsotsos, M. P. Eckstein, and M. S. Landy. Computational models of visual attention. *Vision Research*, 116, Part B:93 – 94, 2015. Computational Models of Visual Attention. 12, 21

[348] J. M. G. Tsui, J. N. Hunter, R. T. Born, and C. C. Pack. The role of V1 surround suppression in MT motion integration. *Journal of Neurophysiology*, 103(6):3123–3138, 2010. 36, 38

[349] A. Turpin, D. J. Lawson, and A. M. McKendrick. Psypad: A platform for visual psychophysics on the ipad. *Journal of Vision*, 14(3), 2014. 44

[350] S. Ullman. Visual routines. *Cognition*, 18:97–159, 1984. 30

[351] S. Ullman. Sequence seeking and counter streams a computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex*, 5:1–11, 1995. 29

[352] S. Ullman. Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Science*, 11(2):58–64, 2007. 31

[353] S. Ullman, L. Assif, E. Fataya, and D. Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences - USA*, 113(10):2744–2749, 2016. 20

[354] S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5(7):682–687, 2002. 31

[355] L. Ungerleider and M. Mishkin. *Two cortical visual systems*, pages 549–586. MIT Press, 1982. 7

[356] L. G. Ungerleider and J. V. Haxby. 'what' and 'where' in the human brain. *Current Opinion in Neurobiology*, 4(2):157–165, 1994. 7

[357] D. R. Valeiras, G. Orchard, S.-H. Ieng, and R. Benosman. Neuromorphic event-based 3d pose estimation. *Frontiers in Neuroscience*, 1(9), Jan. 2016. 26

[358] D. C. Van Essen. Organization of visual areas in macaque and human cerebral cortex. In L. Chapula and J. Werner, editors, *The Visual Neurosciences*. MIT Press, 2003. 7

[359] R. VanRullen and S. J. Thorpe. Surfing a spike wave down the ventral stream. *Vision Research*, 42:2593–2615, 2002. 26, 41

[360] A. Verri and T. Poggio. Against quantitative optical flow. In *Proceedings First International Conference on Computer Vision*, pages 171–180. IEEE Computer Society, 1987. 33

[361] A. Verri and T. Poggio. Motion field and optical flow: qualitative properties. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(5):490–498, 1989. 33

[362] P. Vetter and A. Newen. Varieties of cognitive penetration in visual perception. *Conscious Cognition*, 27:62–75, 2014. 17

[363] W. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287:1273–1276, 2000. 35

[364] R. von der Heydt. Figure-ground organisation and the emergence of proto-objects in the visual cortex. *Frontiers in Psychology*, 6(1695), 2015. 10

[365] C. von der Malsburg. The correlation theory of brain function. Internal report, 81-2, Max-Planck-Institut für Biophysikalische Chemie, 1981. 13

[366] C. Von der Malsburg. The what and why of binding: the modeler's perspective. *Neuron*, 24(1):95–104, 1999. 13

[367] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013. 44

[368] W. Warren. Does this computational theory solve the right problem? Marr, Gibson and the goal of vision. *Perception*, 41(9):1053–1060, 2012. 20

[369] B. Webb, C. Tinsley, N. Barraclough, A. Parker, and A. Derrington. Gain control from beyond the classical receptive field in primate visual cortex. *Visual Neuroscience*, 20(3):221–230, 2003. 13

[370] A. Wedel, D. Cremers, T. Pock, and H. Bischof. Structure- and motion-adaptive regularization for high accuracy optic flow. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1663–1668, 2009. 38

[371] U. Weidenbacher and H. Neumann. Extraction of surface-related features in a recurrent model of V1-V2 interaction. *PLoS ONE*, 4(6), 2009. 32

[372] R. M. Willems. Re-Appreciating the Why of Cognition: 35 Years after Marr and Poggio. *Frontiers in Psychology*, 2, 2011. 17

[373] J. Williford and R. von der Heydt. Border-ownership coding. *Scholarpedia*, 8(10):30040, 2013. 29

[374] A. Witkin and J. Tenenbaum. *Human and Machine Vision*, chapter On the role of structure in vision, pages 481–543. Academic Press, 1983. 28

[375] A. Wohrer and P. Kornprobst. Virtual Retina : A biological retina model and simulator, with contrast gain control. *Journal of Computational Neuroscience*, 26(2):219, 2009. DOI 10.1007/s10827-008-0108-4. 24, 25, 40, 41, 42, 46

[376] J. M. Wolfe, A. Oliva, T. S. Horowitz, S. J. Butcher, and A. Bompas. Segmentation of objects from backgrounds in visual search tasks. *Vision Research*, 42(28):2985 – 3004, 2002. 28, 42

[377] D. Xiao, S. Raiguel, V. Marcar, J. Koenderink, and G. A. Orban. Spatial heterogeneity of inhibitory surrounds in the middle temporal visual area. *Proceedings of the National Academy of Sciences*, 92(24):11303–11306, 1995. 13

[378] N. Yabuta, A. Sawatari, and E. Callaway. Two functional channels from primary visual cortex to dorsal visual cortex. *2001*, 297-301, 292. 11

[379] D. Yamins and J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. 18

[380] D. Yamins, H. Hong, C. Cadieu, E. Solomon, D. Seibert, and J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*, 111(23):8619–8624, 2014. 43

[381] Z. Yang, D. Heeger, B. R., and E. Seidemann. Long-range traveling waves of activity triggered by local dichoptic stimulation in V1 of behaving monkeys. *Journal of Neurophysiology*, 2014. 42

[382] A. L. Yarbus. *Eye movements and vision*, chapter 1-7. Plenum Press, 1967. 32

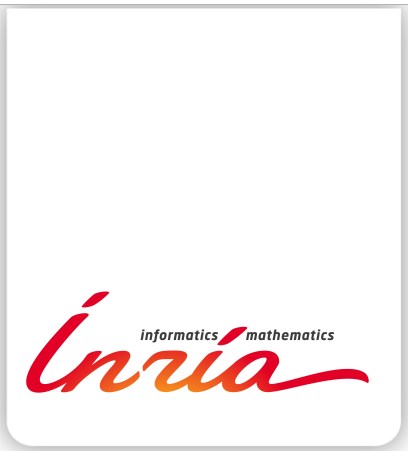
[383] S. Zeki. *A vision of the brain.* Blackwell Scientific Publications, 1993. 9

[384] Y. Zhang, I.-J. Kim, J. R. Sanes, and M. Meister. The most numerous ganglion cell type of the mouse retina is a selective feature detector. *Proceedings of the National Academy of Sciences*, 109(36):E2391–E2398, 2012. 16

[385] H. Zhou, H. S. Friedman, and R. von der Heydt. Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, 20(17):6594–6611, 2000. 10, 29, 42

[386] Y. Zhuo, T. Zhou, H. Rao, J. Wang, M. Meng, M. Chen, C. Zhou, and L. Chen. Contributions of the visual ventral pathway to long-range apparent motion. *Science*, 299(5605):417, 2003. 10

[387] S. Zucker. Computer vision and human perception. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 1102–1116, 1981. 20



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399